APPLICATION OF EXPLAINABLE ARTIFICIAL INTELLIGENCE IN WASTEWATER TREATMENT PLANTS

by

Fuad Bin Nasir

A Dissertation Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in Engineering

at

The University of Wisconsin-Milwaukee

August 2025

ABSTRACT

APPLICATION OF EXPLAINABLE ARTIFICIAL INTELLIGENCE IN WASTEWATER TREATMENT PLANTS

by

Fuad Bin Nasir

The University of Wisconsin-Milwaukee, 2025
Under the Supervision of Professor Jin Li

Wastewater treatment plants (WWTPs) play a crucial role in protecting public health and the environment by removing contaminants before releasing them back into the environment. However, efficient management of these plants can be challenging because of the complex nature of wastewater treatment processes. Machine learning (ML), a subfield of artificial intelligence (AI), can predict WWTP variables by extracting correlations between variables from historical data. Accurate effluent variable prediction through ML can facilitate efficient adjustment of operational variables, thus minimizing operation cost while effectively meeting effluent quality standards. However, relying solely on ML models without understanding the contexts of the predictions is not ideal, especially in WWTPs where operators need to understand the reasons behind model predictions to increase their confidence in real-world applications. While recent studies have focused on predicting variables in WWTP using ML, research on implementing explainable artificial intelligence (XAI) is still developing. XAI is a promising approach for enhancing the transparency and interpretability of ML models in complex environmental systems, such as WWTPs. This study investigated the

performance of widely used ML algorithms in predicting the key effluent quality variables across four WWTPs in Wisconsin, USA. Data was collected from both small- and large-scale WWTPs to evaluate the scalability and generalizability of the models. Feature selection (FS) techniques were used to identify the most relevant features for each target variable in the dataset. XAI tools, including SHapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME), were used to interpret the influence of input variables on the model outputs. This study highlights how ML can accurately predict water quality outcomes, thereby allowing plant operators to better manage their facilities. The study also demonstrates how XAI can clarify the reason behind specific ML models' predictions, enhancing operator confidence in decision-making. These XAI methods identify which factors, i.e., nutrient levels, flow rates, and organic material loads, most significantly impact effluent quality. Using these transparent predictive tools, wastewater facilities can optimize operations, meet regulatory standards, reduce operating costs, and ultimately support healthier communities and ecosystems.

To

my beloved family

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

T	Ton
TPD	Tons per day
min	Minute
F	Fahrenheit
MGD	Million Gallons per Day
lbs/day	Pounds Per Day
SCFM	Standard Cubic Feet Per Minute
CFU	Colony-Forming Unit
MPN	Most Probable Number
MW	Megawatt
KW	Kilowatt
Dth	Dekatherm
MWh	Megawatt Hours
MMCF	Million Cubic Feet

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my advisor Prof. Dr. Jin Li, for her continuous support, guidance, and encouragement throughout my doctoral journey. I am also grateful to my doctoral committee members: Prof. Dr. Qian Liao, Prof. Dr. Yin Wang, Prof. Dr. Zhen Zeng, and Prof. Dr. Xiaoli Ma for their insightful feedback and support, which helped me to improve the quality of my work.

I would like to acknowledge the Department of Civil & Environmental Engineering for providing financial support to pursue my doctoral studies. Thank you to the administrative and support staff for providing the necessary technical and administrative assistance throughout my doctoral study.

I am thankful to my fellow labmates and colleagues for encouraging me throughout my Ph.D. study.

Lastly, I am thankful to my family for their unconditional support, patience, encouragement, and belief in me.

CHAPTER 1: INTRODUCTION

A wastewater treatment plant (WWTP) is an energy-intensive and complex nonlinear process that utilizes physical and microbiological reactions to remove pollutants from wastewater. Increasing urban wastewater and rigorous discharge regulations pose significant challenges for WWTPs to meet regulatory compliance while minimizing operational costs. WWTP's function and effectiveness are significantly influenced by the quality of the treated effluent, and the process must balance energy conservation and pollutant reduction throughout its operation. Thus, predicting effluent concentration is crucial to efficient plant operation and the efficiency of quality control. Traditional methods of variable measurements in WWTP involve time-consuming and labor-intensive sampling and laboratory analysis, high operational costs, and infrequent monitoring. Sensor-based systems require significant investment for installation, calibration, and maintenance. The complex nature of the treatment process, which demands high precision in achieving the desired standard limits of several parameters, is the main barrier to improving effluent quality and meeting the regulatory standard from WWTP while reducing the costs of operation and maintenance.

1.1 Machine Learning and Artificial Intelligence in wastewater treatment plant

WWTPs may improve their decision-making processes, optimize resource allocation, and more successfully handle complicated problems by utilizing the capabilities of machine learning (ML) and artificial intelligence (AI). These systems can dynamically alter process variables, doses, and treatment plans to obtain the best performance. While increasing overall treatment

efficiency, ML based models can reduce operational costs and have a smaller negative impact on the environment. The accuracy and reliability of effluent quality predictions in WWTPs can be enhanced by studying the dynamic interactions between these parameters and applying ML algorithms. Given the complexity of WWTP processes, it is likely that intricate interactions exist among various variables, which have not been extensively explored in earlier research. Additionally, attempts to minimize the number of input parameters might have overlooked certain factors that could have played crucial roles in specific contexts.

1.2 Understanding ML models

ML algorithms are traditionally black-box in nature. ML models without understanding the contexts of predictions are not ideal, especially in WWTPs, where operators need to understand the reasons behind model predictions to increase their confidence in real-world applications. Explainable artificial intelligence (XAI) is capable of making black box models easier to understand. While recent studies have focused on predicting variables in WWTP using ML, research on implementing XAI is still developing. Moreover, a comparative study that assesses ML performance in both small-scale and large-scale WWTPs has not been thoroughly explored.

1.3 Objective of the dissertation

The main goal of the dissertation is to bridge the existing gap between ML model interpretability and the accurate prediction of water quality variables. The objectives can be summarized as follows.

Firstly, investigating the use of ML models for WWTP sludge predictions and XAI techniques for understanding the impact of variables behind the prediction. This study provides an example of using ML models in sludge production prediction and applying XAI to understand factors influencing it. An understandable interpretation of ML model predictions can facilitate targeted interventions for process optimization and improve the efficiency and sustainability of wastewater treatment processes. The purpose of the study is to use ML models in sludge production prediction and interpret models applying XAI to understand factors influencing it.

Secondly, exploring the application of several ML models specifically, Artificial Neural Networks (ANN), Gradient Boosting Machines (GBM), Random Forests (RF), eXtreme Gradient Boosting (XGBoost), and hybrid RF-GBM models in predicting important WWTP variables such as Biochemical Oxygen Demand (BOD), Total Suspended Solids (TSS), Ammonia (NH_3), and Phosphorus (P). The significance of the study lies in investigating several ML models' performance in predicting multiple influential WWTP variables.

Finally, investigating the performance of two widely used ML algorithms, XGBoost and RF, in predicting the key effluent quality variables across four WWTPs in Wisconsin, USA. The target effluent variables included ammonia nitrogen (NH_3-N), BOD, chemical oxygen demand (COD), total phosphorus (TP), and TSS. The purpose of the study is to investigate the performance of two widely used ML algorithms in predicting the key effluent quality variables across several WWTPs.

CHAPTER 2: UNDERSTANDING MACHINE LEARNING PREDICTIONS OF WASTEWATER TREATMENT PLANT SLUDGE WITH EXPLAINABLE ARTIFICIAL INTELLIGENCE

2.1 Introduction

Sludge management in wastewater treatment plants (WWTPs) is a complex and dynamic process, influenced by several factors, including treatment technology selection, operational expertise, and influent characteristics (Ekinci et al., 2023). Accurate sludge production prediction plays a crucial role in optimizing WWTP operation and minimizing costs associated with sludge handling and disposal. Existing methods for sludge prediction often rely on traditional techniques with constraints such as limited accuracy, inflexibility, and difficulty in capturing complex relationships between variables.

Machine learning (ML), a subset of artificial intelligence (AI), is becoming popular in predicting variables in WWTP because ML algorithms can handle complex nonlinear, unstable, and interdisciplinary features without any expert knowledge (Bagherzadeh et al., 2021; Ly et al., 2022; Tung & Yaseen, 2020; Xu et al., 2022). The use of ML in the WWTP resulted in a significant reduction of operational costs (Wang et al., 2024). A number of studies have been carried out using various ML models to predict the influent and effluent characteristics of WWTPs (Guo et al., 2015; Nourani et al., 2018; Wang et al., 2021; El-Rawy et al., 2021; Azimi et al., 2022; Ching et al., 2022; Safder et al., 2022; Li et al., 2022; Zhu et al., 2022; Yadav et al., 2023; Yu et al., 2023; Duarte et al., 2023). While sludge management is a critical aspect of WWTP operation, only a handful of research articles used ML models to predict sludge (Ekinci et al., 2023; Hu et al., 2024; Shao et al., 2023; Zeinolabedini & Najafzadeh, 2019). Ekinci et al. (2023) achieved a significant level of accuracy in sludge prediction. However, the study

considered only a limited set of data that was not representative of different treatment plants with different configurations and larger datasets. While ML models offer a promising approach for accurate sludge prediction, their black-box nature can hinder trust and limit their practical application. Unlike traditional statistical methods that assess variable significance, ML models prioritize prediction accuracy. Less information is given on which variables are truly driving the model's predictions and to what extent. The inability to explain variable importance poses significant challenges that limit trust in the model's results and hinder the ability to improve the model. As a result, WWTP operators still rely on traditional methods of sludge prediction, such as estimation based on historical sludge generation, which neglects influent variation as well as internal correlation between variables.

There is a need to bridge the gap between ML research and industrial practice in sludge management (Hu et al., 2024). Recently, there has been a noticeable transition from black box to explainable ML models, particularly in fields where experts expect accurate models with an explanation of the generated outcomes, thereby increasing confidence in model effectiveness for real-world applications (Park et al., 2022). Recent advancements in the field of explainable artificial intelligence (XAI) aim to improve the interpretability of black box models. The application of XAI in sludge production prediction can help better explain the prediction results. Concerns such as the rationale behind model predictions, the basis for trust in these predictions, and strategies for error correction are particularly relevant in the WWTP. Moreover, the complexity of sludge production mechanisms and the difficulties encountered in field monitoring result in limited data availability for research. A comprehensive set of data provides a better understanding of the contribution of a wide range of variables in sludge

production. The integration of XAI tackles constraints associated with the black-box nature of models, where the decision-making process of the model is not transparent.

No dedicated research has been published so far that uses XAI in predicting sludge generation for US WWTPs. Hu et al. (2024) have explored similar concepts within Chinese facilities and encountered limitations in accuracy and variable consideration. The present study aims to address this knowledge gap by applying XAI methods to enhance the interpretability of ML models for predicting sludge in a WWTP in Milwaukee, Wisconsin, USA. In 2022, Wongburi and Park (2022) studied the prediction of the sludge volume index of a US WWTP. However, the broader scope of sludge prediction was not addressed by incorporating multiple XAI approaches, nor by employing various feature selection (FS) and ML methods to compare performances. Our objective in the study is to enhance confidence and transparency while gaining valuable insights by integrating XAI. The transparency allows for a deeper understanding of the model's decision-making process and facilitates informed decision-making based on both data and human expertise. XAI techniques can reveal the most influential factors contributing to sludge production and thus offer essential insights for optimizing processes. The knowledge gained can help to identify areas for improvement and implement targeted interventions to minimize sludge generation, leading to cost savings and environmental benefits.

In this study, data for 73 water quality, water quantity, and electrical variables were collected from a WWTP, and several FS methods were used to reduce the number of input variables. Using various ML models, we focused on predicting primary sludge and waste activated sludge (WAS) because they are the primary components of sludge production in WWTP and play

the most significant role in management and operations. XAI methods were used to understand the reason behind the ML model's prediction. XAI and FS techniques were able to understand the transparency and fairness in AI-driven decisions. The findings will facilitate more reliable, transparent, and data-driven approaches to sludge prediction, ultimately enhancing the efficiency and sustainability of wastewater treatment processes.

2.2 Methods

Data collection, data preprocessing, and data prediction are the three key elements for evaluating and predicting the target variable of a WWTP. After collecting and preprocessing data, feature selection was conducted to identify and choose pertinent features that were crucial for predicting the target variable.

2.2.1 Data collection

The data were collected from Southshore WWTP in Milwaukee, Wisconsin, USA. The facility treats wastewater from industrial, municipal, and domestic sources, with a maximum daily flow of 250 million gallons per day (MGD) and a peak hourly flow of 300 MGD. The facility includes influent flow monitoring, mechanical bar screens, grit chambers, primary clarifiers, aeration basins, secondary clarifiers, disinfection, and effluent pumping. The water quality, quantity, and electrical data from various locations, including influent, primary influent, primary effluent, aeration basins, secondary clarifier, secondary effluent, effluent, blowers, and lab of the facility, were collected. Initially, 73 daily and hourly variables starting from January 1, 2019, to January 28, 2024 (Hourly variables were converted to daily values) were collected. After data processing, 57 variables with 105,687 entries were considered in the study. Table 2.1 presents

the list and descriptions of variables. The distribution of primary sludge and WAS data is presented in Figure 2.1. The distribution of the 57 variables and a list of abbreviations is presented in the supporting information.

2.2.2 Data preprocessing

Data obtained by the sensors typically contain abnormalities related to recording, such as the occurrence of extreme values. The obtained dataset was examined for missing or incorrect values, and a few abnormal values were found in “Detention time” and “MCRT (EP + WP) Days.” Examples of abnormal values include having “MCRT (EP + WP) Days” values as 60790.59, 74038.83, 122488.83, 69090.51, 2255.59, and 67828.00 or “Detention time” as high as 90,000 min. Abnormal values were detected by human observation and replaced with an average value of the variable. The dataset also contained some missing values which were replaced with an average value of the variable. Outliers were kept in the dataset in order to understand the whole picture of the analysis as suggested by Ly et al. (2022). To reduce multicollinearity, variables that have high correlation (greater than or equal to|0.9|) were removed.

2.2.3 Feature Selection

Traditional ML predictions frequently include many input variables, which can be expensive computationally and difficult to monitor in real time. The possibility of reducing the input variable set while keeping high prediction accuracy were investigated in the study to allow more efficient and economical control of WWTPs. Usually, FS techniques are utilized during data preprocessing steps to enhance the accuracy of classification or regression tasks,

wherein features are prioritized based on their significance to the target variable (Ekinci et al., 2023). The most important variables for predicting sludge were identified using some well-known FS approaches, including least absolute shrinkage, and selection operator (LASSO), mutual information (MI), random forest (RF), and Pearson correlation (PC) as described by Bagherzadeh et al. (2021) and Xu et al. (2024). LASSO scores range from negative to positive values while PC scores range from -1 to 1. MI score varies from 0 (no information shared) to positive values, and RF produces a feature importance score that ranges from 0 to 1, with a score of 0 suggesting a feature that was not used in the prediction and a score of 1 indicating a feature that contributed to the output's perfect prediction. LASSO was chosen as it can identify the most important features and is effective in dealing with datasets with many features. In contrast to the correlation coefficient, which is limited to capturing linear dependencies between variables, MI can detect both linear and nonlinear relationships, a characteristic that has driven its popularity in FS tasks. The RF model was used to select input parameters because it generates robust variable importance ratings, effectively handles multicollinearity, and captures nonlinear connections and interactions between parameters (Tyralis et al., 2019). PC was initially used to find out the linear dependence between each pair of variables and to discard strongly correlated variables as suggested by Xu et al. (2024). Then PC was used again to select a reduced number of input variables.

Table 2.1. Description of variables

Variable	Average	Standard Deviation	Missing rate
$Flow_i$ (MGD)	89.34	39.23	0.00
$(NH_3)_i$(mg/L)	22.29	7.72	0.00
BOD_i (mg/L)	330.82	155.02	0.49
TSS_i (mg/L)	252.02	104.13	0.00
$(NH_3)_e$ (mg/L)	0.94	1.55	0.00
BOD_e (mg/L)	1300	6.66	0.43
P_e (mg/L)	0.47	0.32	0.00
TSS_e (mg/L)	9.26	6.18	0.05
TSS_i Load (T)	86.78	34.07	0.00
BOD_i Load (T)	108.76	36.8	0.54
$(NH_3)_i$ Load (T)	7.26	1.33	0.00
P_i Load (T)	4593.09	1301.24	0.00
Primary sludge (TPD)	54.04	20.65	0.32
TSS_{pe} (mg/L)	107.44	104.33	2.40
Iron Dose (mg/L)	10.98	4.68	17.75
Ferric dose (mg/L as Fe)	4.59	6.49	0.00
WP Aer Basin Mass (TPD)	190.75	42.54	0.97
EP Aer Basin Mass (TPD)	191.02	40.89	1.61
Aerobic MCRT (Days)	7.99	2.02	2.59
EP Plant MLSS (mg/L)	2939.67	480.71	1.56
WP Plant MLSS (mg/L)	2911.07	460.19	0.86
Ortho P_e (mg/L)	0.30	0.28	0.81
Detention Time (min)	94.38	34.55	0.00
Aer Basin Temp (F)	59.6	5.74	1.02
DO Set Pt (mg/L)	3.63	0.36	1.67
WP Avg DO (mg/L)	3.64	1.47	1.73
EP Avg DO (mg/L)	3.22	1.20	11.11
SVI (mL/g)	113.89	39.37	0.00
MCRT (Days)	10.28	2.60	1.13
WAS (TPD)	37.86	13.16	1.61
WAS Flow (MGD)	1.61	0.45	0.00
DSD Mass to JI (TPD)	45.01	8.86	2.86
EP $(NO_2^-)_{se}$	0.27	0.28	2.10
WP $(NO_2^-)_{se}$	0.28	0.29	1.56
WP $(NO_3\text{-})_{se}$	3.63	1.99	1.56
EP $(NO_3\text{-})_{se}$	3.64	2.00	2.10
WP $(NH_3)_{se}$	0.92	1.66	1.51
EP $(NH_3)_{se}$	0.84	1.59	2.04
pH_e	7.18	0.09	0.05

TSS_e Removed (%)	95.68	3.85	0.00
BOD_e Removed (%)	95.14	4.34	0.54
$Temp_e$ (F)	157.00	83.57	0.00
TRC (mg/L)	0.01	0.01	0.00
GBT Polymer Used (lbs/day)	5074.34	5658.02	8.19
Total DG Production (MMCF)	1.63	0.63	0.00
Total Dig Gas Flow (SCFM)	1109.96	288.93	0.38
Digester MCRT (Days)	26.96	9.09	3.02
Fecal Coliforms (CFU/100ml)	304.35	1804.94	0.43
E coli (MPN/100ml)	1008.57	8668.71	0.20
Total Plant Load (MW)	5.21	0.50	0.00
Total Elec Generated (MW)	3.34	0.82	0.00
Total Blower Elec Used (KW)	2858.84	331.09	0.00
Total DG used (Dth)	789.83	202.48	0.05
Elec Generated from DG (MW)	63.40	18.45	0.05
Elect Generated from NG (MWh)	16.76	17.74	0.05
Methane DG (%)	58.76	2.10	0.00
Dig Gas to Flares (MMCF)	0.33	0.35	0.22

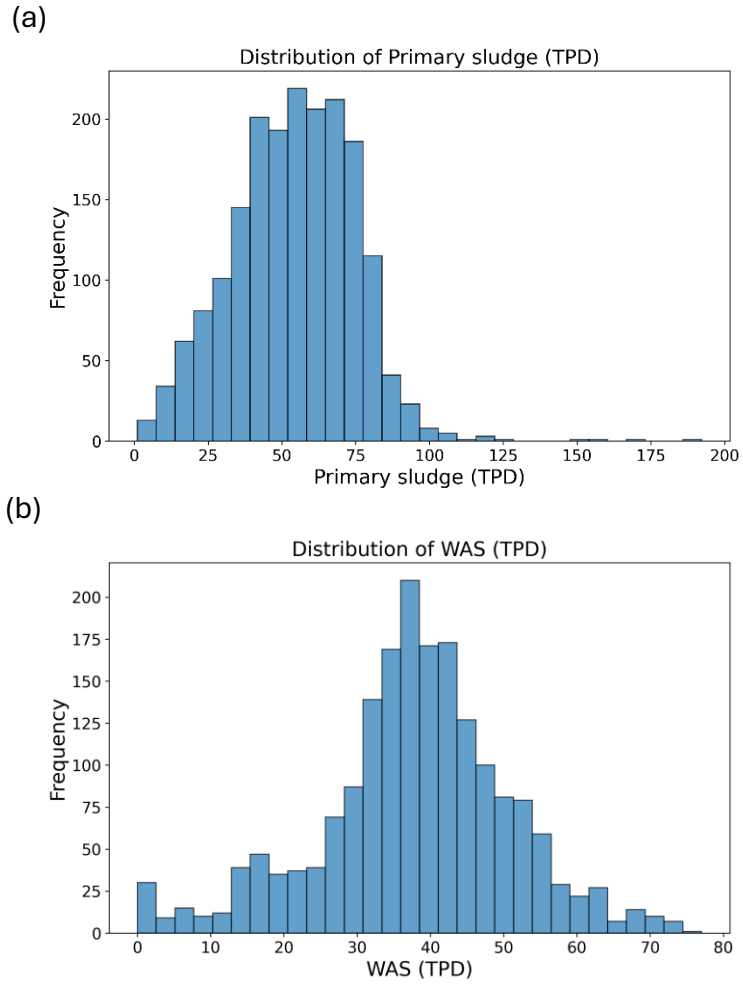

Figure 2.1. Distribution of Primary sludge (a) and WAS (b)

2.2.4 SHapley additive exPlanations (SHAP)

SHapley additive exPlanations (SHAP) were also used to obtain a reduced set of variables. In addition to the full set of data, the most significant 10, 5, 3, and 2 variables selected by the FS method were considered as input variable combinations. While many of the most influential variables related to primary sludge and WAS were shared by the selection methods, their importance hierarchy was different. Therefore, several input variable combinations were chosen for the study. Because the total sludge in the facility is the combination of primary

sludge and WAS, both primary sludge and WAS were predicted. Shapley values from game theory to attribute ϕ_i values to each feature are explained as follows (Lundberg et al., 2018; Lundberg & Lee, 2017; Xu et al., 2024):

$$\phi_i = \sum_{S \subseteq F/\{i\}} \frac{|S|!\,(|F|) - |S| - 1)!}{|F|!} [f_{SU\{i\}}(x_{SU\{i\}}) - f_S(x_S)] \qquad 1$$

Where ϕ_i is the SHAP of the *i*th input variable, F is the set of all input variables, S is the subset of all input variables used in the model, $f_{SU\{i\}}$ is trained with that feature present, and another model f_S is trained with the feature withheld, x_S represents the values of the input variables in the set S. whereas $x_{SU\{i\}}$ represents the data set that includes the *i*th input variable.

2.2.5 Local interpretable model-agnostic explanation (LIME)

LIME is an XAI tool that interprets black-box ML model with a local, interpretable model to explain each prediction (Hu et al., 2024). It identifies the top features that contribute the most to the prediction made by the model. Each feature is associated with a weight that indicates its impact on the prediction. Features with positive weights have a positive impact on the prediction, while features with negative weights have a negative impact. The magnitude of the weight indicates the strength of the feature's influence on the prediction. The features are sorted based on their importance, with the most influential features appearing at the top. The explanation produced by LIME is obtained by the following (Ribeiro et al., 2016):

$$\xi\,(x) = \underset{g \in G}{\operatorname{argmin}}\ \mathcal{L}(f, g, \Pi_x) + \Omega\,(g) \qquad 2$$

Where, $\mathcal{L}(f, g, \Pi_x)$ is the measure of how unfaithful g is in approximating f in the locality defined by Π_x, $\mathcal{L}$ is the fidelity function, G is the class of potentially interpretable models, and Ω is the complexity measure. LIME experiences a tradeoff between model fidelity and complexity.

2.2.6 ML models

To predict sludge, widely used ML models, random forest (RF), gradient boosting machine (GBM), and gradient boosting tree (GBT) were used in the study. RF is an ensemble learning technique and one of the most influential ML methods. It uses several decision trees to produce predictions (Jiang et al., 2023; Sun et al., 2024; Szomolányi & Clement, 2023; Tyralis et al., 2019) and is the only algorithm that provides insights into the importance of each variable, a crucial asset for further analysis (Zhang et al., 2021). A boosting approach called GBM combines several weak prediction models, typically decision trees, to produce a powerful predictive model (Konstantinov & Utkin, 2021). To fix the errors created by the previous trees, GBM iteratively adds new models. Like GBM, GBT makes predictions using an ensemble of decision trees. However, GBT reduces the model complexity by optimizing the tree building using gradient descent and a regularization term. We employed the GradientBoostingRegressor class from the scikit-learn library to implement the GBT model. This model was chosen because of its robustness, flexibility, and ability to handle complex datasets with nonlinear relationships. XGBoost (Extreme Gradient Boosting) is an optimized and highly efficient implementation of the gradient boosting algorithm. It is known for its superior performance and scalability, making it a popular choice for regression and classification tasks. XGBoost improves upon traditional gradient boosting by incorporating regularization techniques and parallel processing capabilities. We utilized the XGBRegressor class from the XGBoost library to implement the GBM model. Our

selection of RF, GBM, and GBT was motivated by their hybrid nature as they effectively combine the strengths of diverse algorithmic frameworks with the interpretability and structure provided by decision tree structures. These algorithms have demonstrated effectiveness in predicting various variables associated with WWTP (Bagherzadeh et al., 2021; Ching et al., 2022; Hu et al., 2024; Ly et al., 2022; Shao et al., 2023; Sun et al., 2023; Xu et al., 2024; Wei et al., 2023). The present study explored the efficiency of above-mentioned models in WWTP for a large and wide range of datasets.

2.2.7 Model training and evaluation

In the analysis, the dataset was divided into a training set and a test set. Splitting the data ensures that the models are trained on a representative subset of the data and evaluated on test (unseen) data, giving a trustworthy assessment of their generalization ability (Yadav et al., 2023). The datasets were divided following the typical ratios between testing and training data according to the literature (Ly et al., 2022; Xie et al., 2022; Xu et al., 2024). Out of the 1854 days of data, 90:10, 80:20, and 70:30 ratios were considered as training and test datasets. Different sets of training and test data were chosen to evaluate the model's generalization ability. Validation is considered a crucial step of the model development process to make sure that the developed model is accurate enough for their intended use (Nourani et al., 2018). Therefore, k-fold cross-validation was also performed to lower the risk of overfitting (Zhang & Liu, 2023) by splitting the entire dataset into five equal sized sections. In each iteration, a different fold was used as the test set, and the remaining folds were used as the training set. The model's performance is influenced by the hyperparameter values used for training. The tuning process for hyperparameters involves experimenting with different configurations for

each hyperparameter to attain the best model performance. Using a well-known hyperparameters tuning method, grid search, various hyperparameter combinations for RF, GBM, and GBT models were tested to find the optimum set that yields the best performance on certain validation metrics. To evaluate the regression model's performance, several model metrics can be used depending on the specific tasks, data characteristics, and circumstances. In this regression study, four widely used assessment metrics - mean absolute error (MAE), mean squared error (MSE), R-squared (R^2), and root mean squared error (RMSE) - were used to evaluate the performance of the ML models. MAE measures the average magnitude of the errors between predicted and actual values (Equation 3). It helps to understand the average deviation of the predictions from the actual values. MSE measures the average squared difference between the predicted values and the actual values (Equation 4). It helps to understand the potential magnitude of predicted deviations. R^2 quantifies the percentage of variance that is explained by the models (Equation 5). R^2 value ranges from 0 to 1, where R^2 value of 1 is perfectly able to explain the variability in the data, and one with a value of 0 suggests that it is unable to explain the variability. RMSE denotes the average size of the residuals (Equation 6). RMSE is always non-negative, and a lower value indicates better fit. These metrics reveal information about the produced ML models' precision, goodness-of-fit, and accuracy.

$$MAE = \frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{n} \qquad 3$$

$$MSE = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n} \qquad 4$$

$$R^2 = 1 - \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}} \qquad 5$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \qquad 6$$

where $\hat{y}_i$ is the predicted value and y_i is the experimental data and *n* is the number of test observations. In our study, we used multiple metrices to gain insight into both the average magnitude and distribution of errors as well as the model's ability to explain the data's variance. Paired t-test was conducted to determine the statistical significance of the differences between training and testing results for each metric of ML models. If there is a statistically significant difference between training and testing results, it suggests that the model might be overfitting to the training data. Overfitting occurs when the model captures noise in the training data, leading to poor generalization of unseen data. If there is no statistically significant difference in MAE, MSE, and RMSE between models for both training and test data, it suggests that, on average, the models are making similar magnitude errors when predicting both training and test data. No statistically significant difference in R^2 between training and test data suggests that the models are explaining a similar proportion of the variance in the data.

2.2.8 Comparison of actual and predicted Data

During both training and testing phases, the actual data on sludge production were compared with the predicted data from the models. This comparison was conducted to ensure that the models could accurately replicate real world observations. The predicted values from the models were validated against the actual sludge production data to assess the models' (RF, GBM, and GBT) performance. By comparing the predicted and actual data, we were able to

evaluate the models' ability to generalize to new, unseen data and ensure their robustness and reliability for predicting sludge production in WWTPs.

2.3 Results and Discussion

2.3.1 Feature selection

A heatmap is a graphical representation of data where the individual values contained in a matrix are represented as colors. In the context of WWTP, a heatmap can provide valuable insights into the relationships and interactions between various variables. A correlation heatmap analysis of WWTP variables revealed interdependence in some variables (Figure 2.2). For instance, strong positive correlations between variables like BOD removed and TSS removed (0.85) underscore the interdependence between the removal efficiencies of BOD and TSS in the effluent.

The top 10 strongly correlated variables related to primary sludge and WAS according to four FS methods are shown in Figure 2.3. Total DG production and effluent temperature repeatedly appeared as key factors affecting primary sludge generation in multiple FS methods. While variables like MCRT and flow consistently emerged as key predictors of WAS production, the consistent identification of these variables across multiple selection methods highlighted their substantial influence on both sludge types based on FS methods. While many of the top 10 most influential variables related to primary sludge and WAS were shared by the selection methods, their importance hierarchy was different. Therefore, in addition to the full set of data consisting of 57 variables, the top 10, 5, 3, and 2 variables were selected as input variables based on FS methods. Moreover, selecting a reduced number of input variables based on FS

methods helps in decreasing computational costs and reducing the risk of overfitting. By focusing on the most impactful variables, the models can deliver high predictive accuracy while remaining computationally efficient.

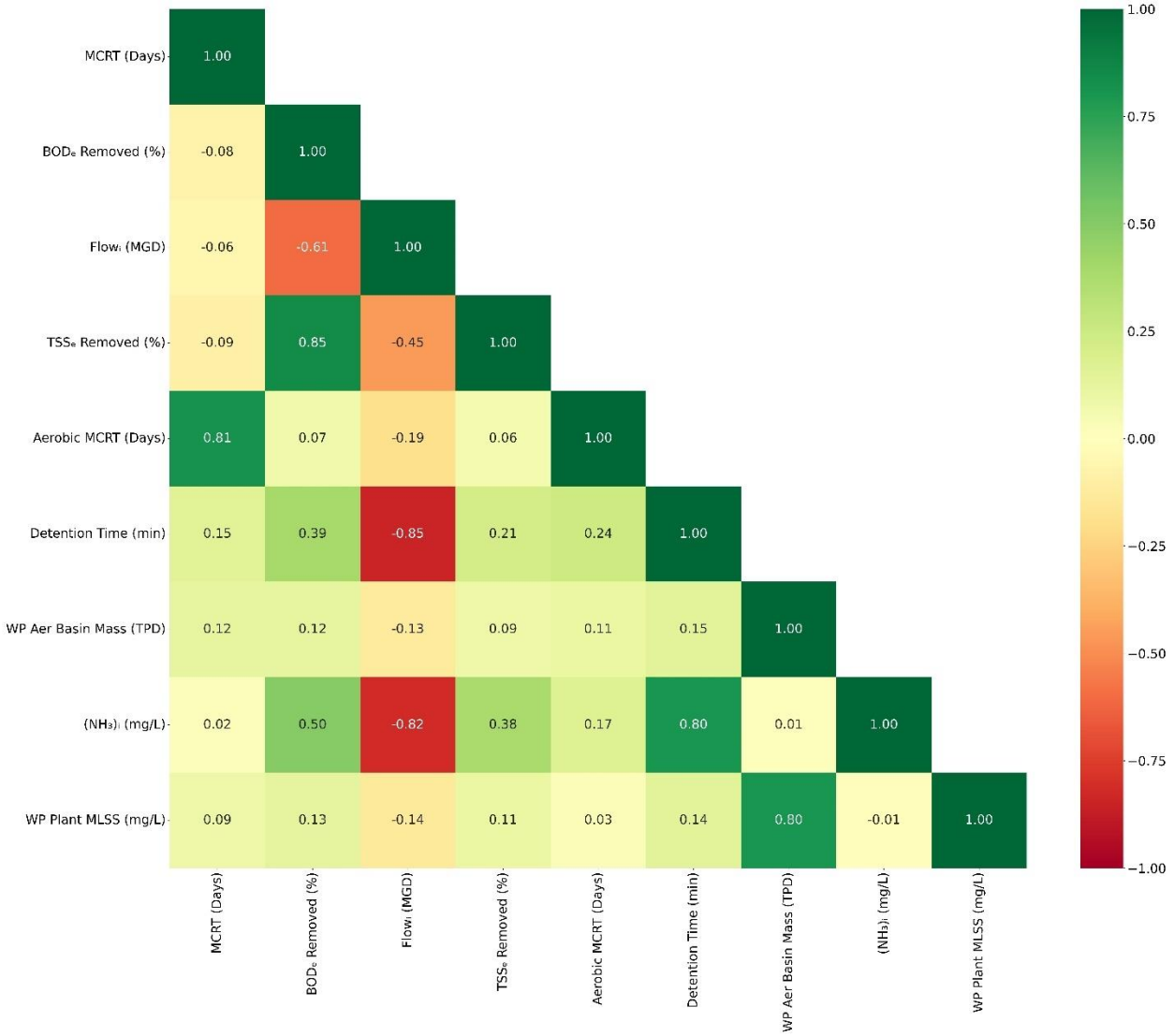

Figure 2.2. Heatmap of the variables with a high correlation (≥|0.8|)

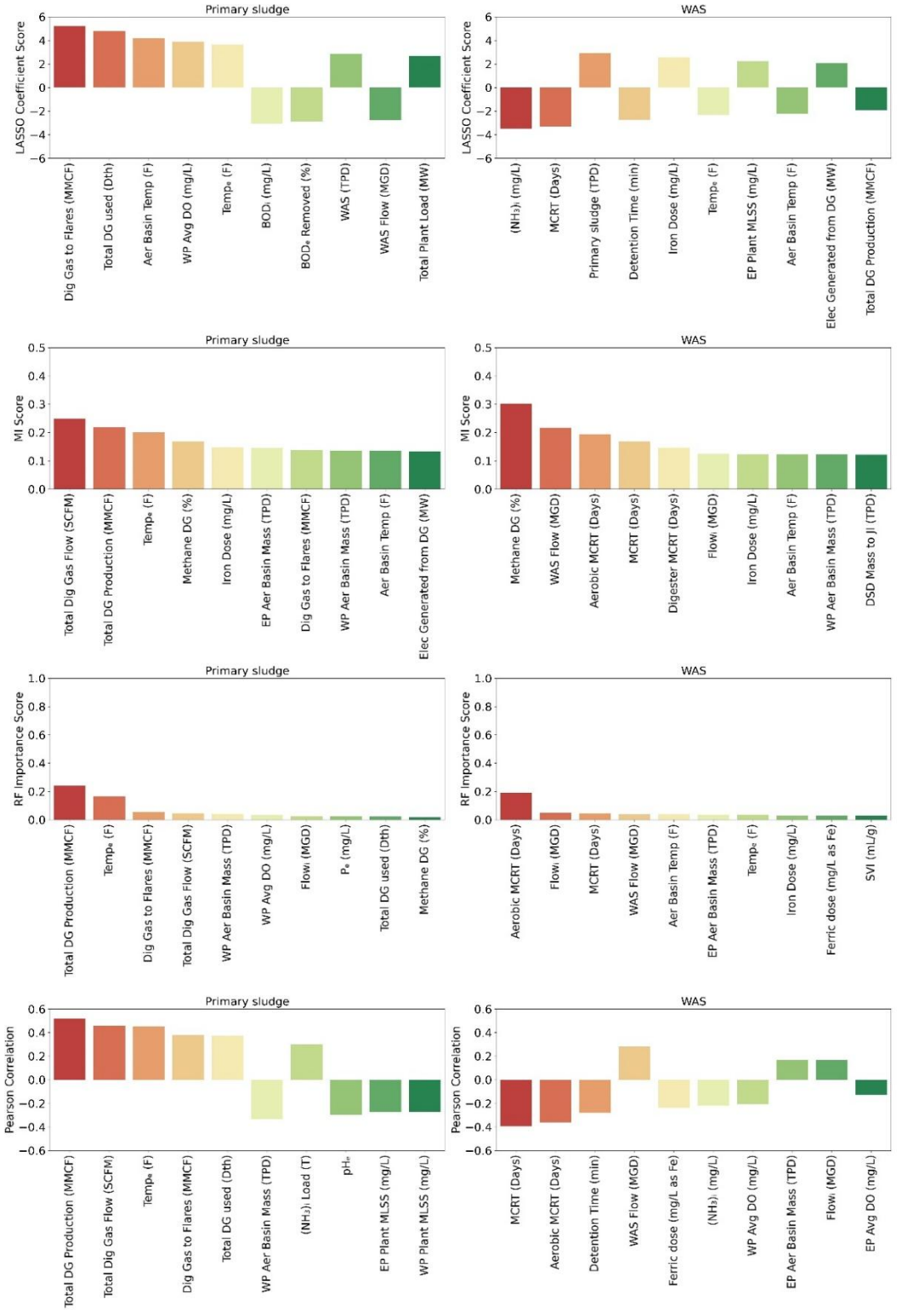

Figure 2.3. Top ten strongly correlated variables related to primary sludge (left) and WAS (right).

2.3.2 SHAP

In Figure 2.4, SHAP summary plots show the relative impact of influent variables on GBM model performance. SHAP values in a higher position indicate greater importance of the input variables on model performance (Park et al., 2022). Red colored dots on the right side represent positive values of SHAP, and blue-colored dots on the left side represent negative values of SHAP. A positive weight indicates that an increase in the value of the feature tends to increase the prediction made by the model. Conversely, a negative weight suggests that an increase in the value of the feature tends to decrease the prediction made by the model. Thus, the feature has a negative influence on the prediction outcome. In predicting primary sludge, RF, GBM, and GBT models' performance increased with higher value of effluent temperature, total DG production, total DG used, average DO, methane DG, WAS, and flow. Thus, according to SHAP value, primary sludge production will increase with the increase of above-mentioned variable. However, higher values of aeration basin mass, P_e, and WAS flow decreased model prediction value. When predicting primary sludge, it was observed that temperature had the most significant impact with all the models. This finding is consistent with existing literature, which emphasizes the significant role of temperature in WWTP (Arnell et al., 2021). In predicting WAS, it was observed that aerobic MCRT had the most significant impact on all the model outputs. Longer MCRT resulted in more degradation of organic matter, less sludge buildup, and improved sludge settling, ultimately reducing the amount of WAS production. Flow, aeration basin mass, digestor MCRT, and WAS flow all made positive impacts on model outputs. Aerobic MCRT, MCRT, aeration basin temperature, GBT polymer used, $(NH_3)_i$, and ferric dose reduced the model output.

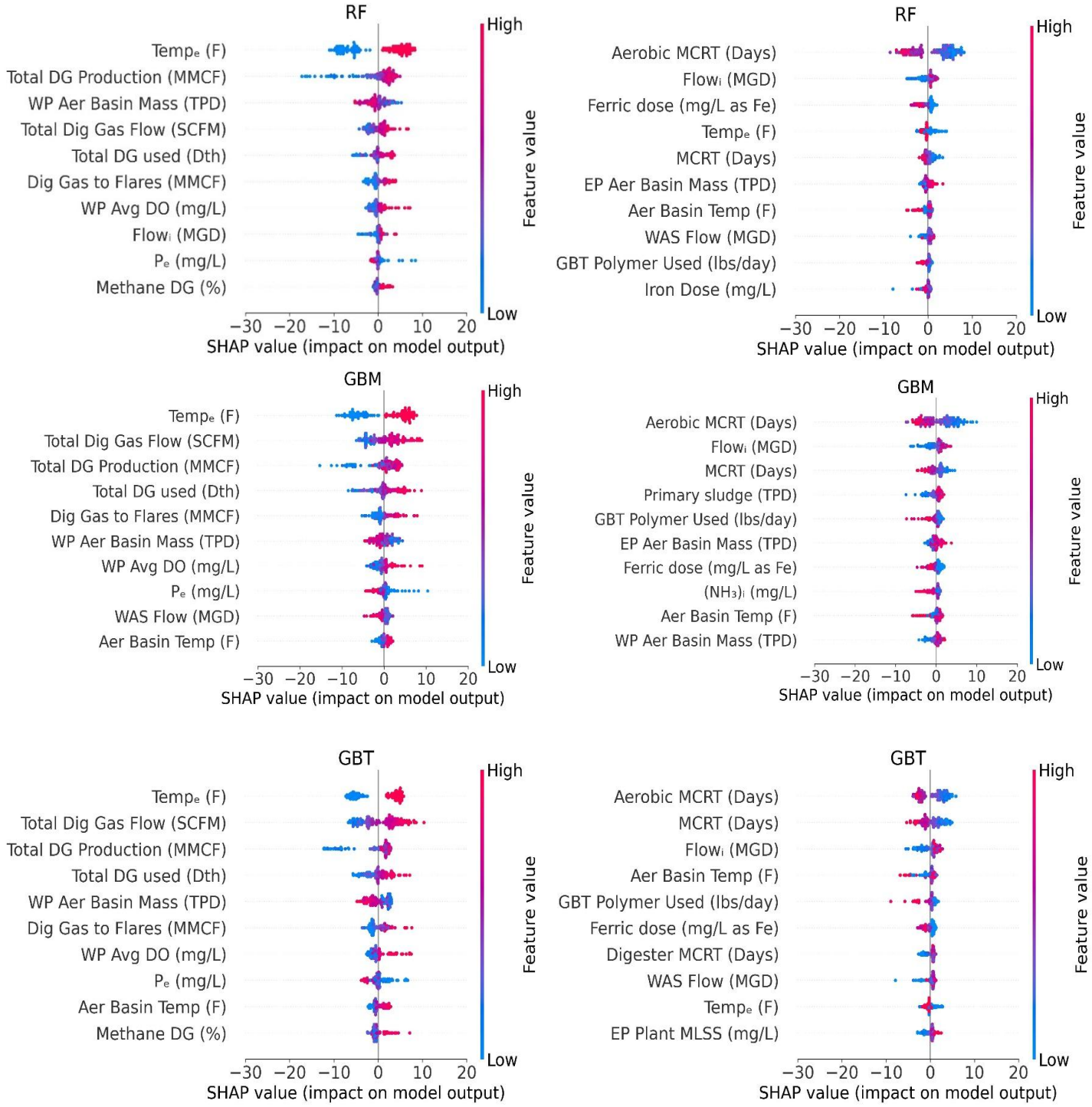

Figure 2.4. SHAP summary plot. Primary sludge (left); WAS (right)

2.3.3 LIME

LIME explains the impact of target variables for specific instances. The instance represents a single data point or observation for which the user can understand the model's prediction and the factors influencing it. The LIME explanation for primary sludge and WAS prediction for GBM model is shown in Figure 2.5. According to LIME, for RF, GBM, and GBT models, the primary sludge productions for the 100th test instance were expected to be around 53.31 TPD, 50.10 TPD, and 55.71 TPD, respectively. For the RF model, the LIME explanation suggested that temperature was the most significant driver. Temperatures between 61.45°F and 228.27°F appeared to increase the predicted value of primary sludge. Other significant drivers such as digestor production, dig gas to flares, and total dig gas flow positively impacted sludge production. Both temperature and digestor gas were the top two significant drivers based on SHAP and LIME. For both GBM and GBT, digestor gas flow took the most significant driver and impacted sludge production positively. While digestor gas to flare and temperature were positive drivers, phosphorus appeared to decrease sludge production. The predicted values of the WAS for the specific instance being explained were approximately 32.62 TPD, 35.90 TPD, and 31.44 TPD for RF, GBM, and GBT respectively.

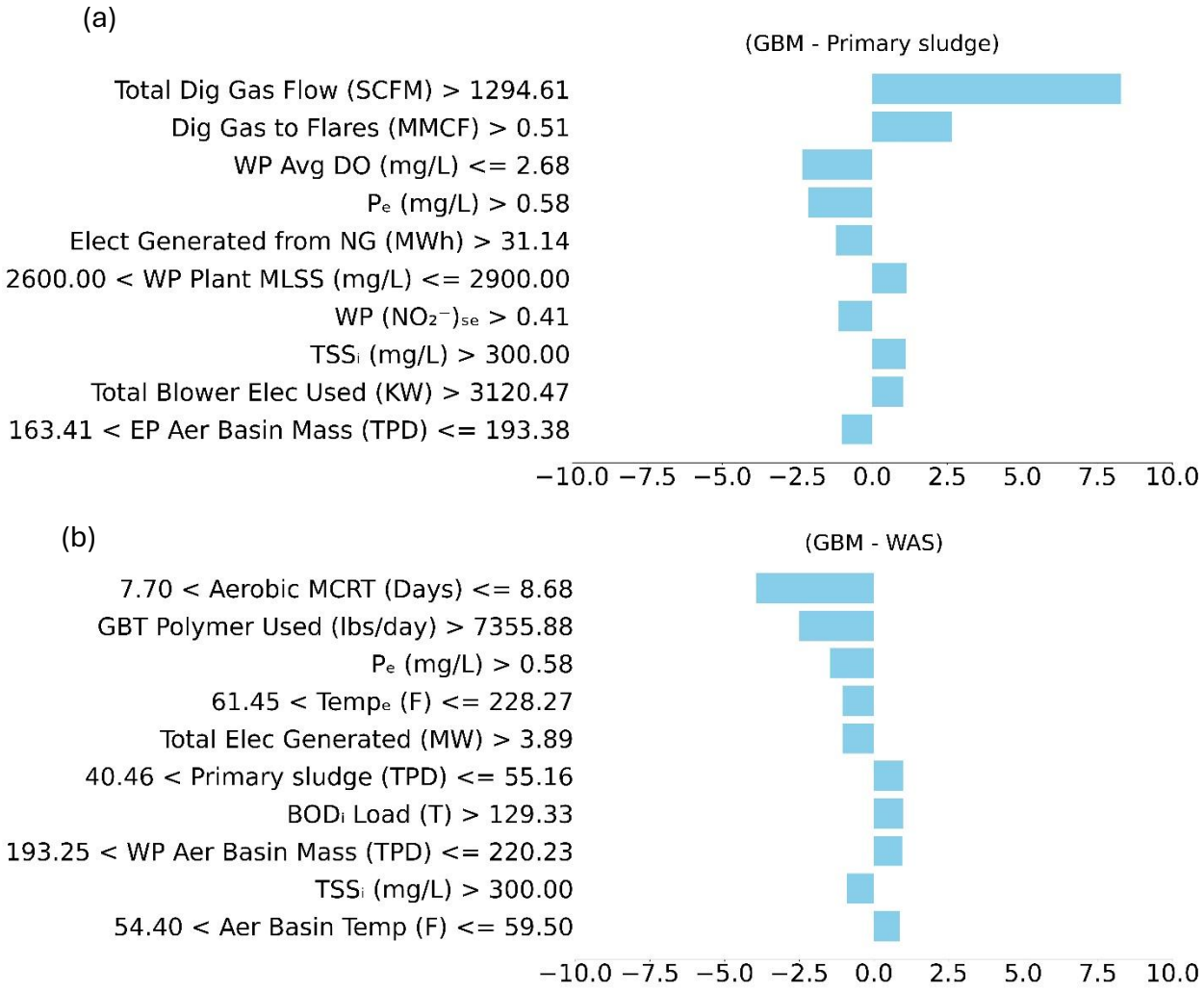

Figure 2.5. LIME explanation for prediction. (a) Primary sludge (b) WAS.

When the aerobic MCRT was greater than 7.70 days and less than 8.68 days, it lowered the predicted value of WAS for all the models. Longer MCRT might indicate higher efficiency in the aerobic treatment process, leading to a lower WAS production. For both RF and GBM, aerobic MCRT stood as the most influential variable according to LIME while for GBT it took second position. Aerobic MCRT was the most influential in SHAP for all models. Temperatures between 61.45°F and 228.27°F tended to decrease the predicted value of WAS. Lower temperatures might slow down biological processes, affecting the generation of WAS. GBT polymer used decreased model prediction value and stayed as the second and third most influential driver in RF and GBM, respectively, and topmost in GBT.

2.3.3 Model performance

The model performance (shown in Tables 2.2) revealed that regardless of FS methods, all models tended to increase R^2 and decrease MAE, MSE, and RMSE with the increase of input variables for both training data and test data. In predicting primary sludge, RF, GBM, and GBT models were able to achieve R^2 values of 0.95, 0.98, and 0.98 for the training dataset and 0.70, 0.70, and 0.71 for the test dataset, respectively. Figures 2.6 and Figure 2.7 show comparison of actual data on sludge production from the systems with the predicted data obtained from the GBM model for full dataset during the training and testing for primary sludge and WAS, respectively. In the study, all the models' performance improved with an increase in input variables. The result agrees with a study on the impact of various FS methods on ML algorithm performance in a WWTP conducted by Bagherzadeh et al. (2021) which reported that adding more features increased accuracy.

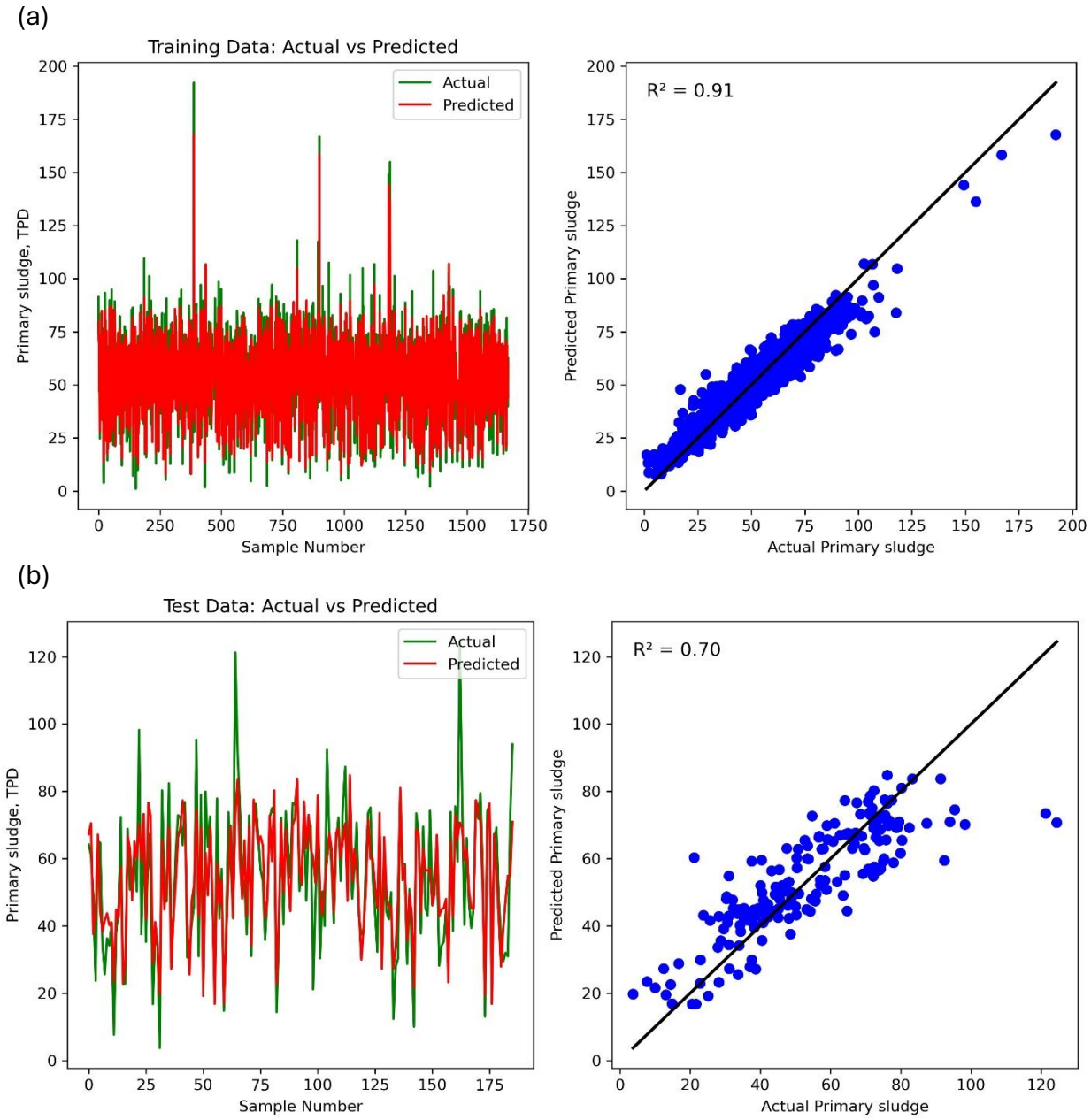

Figure 2.6. GBM model performance in predicting primary sludge. (a) training data (b) test data

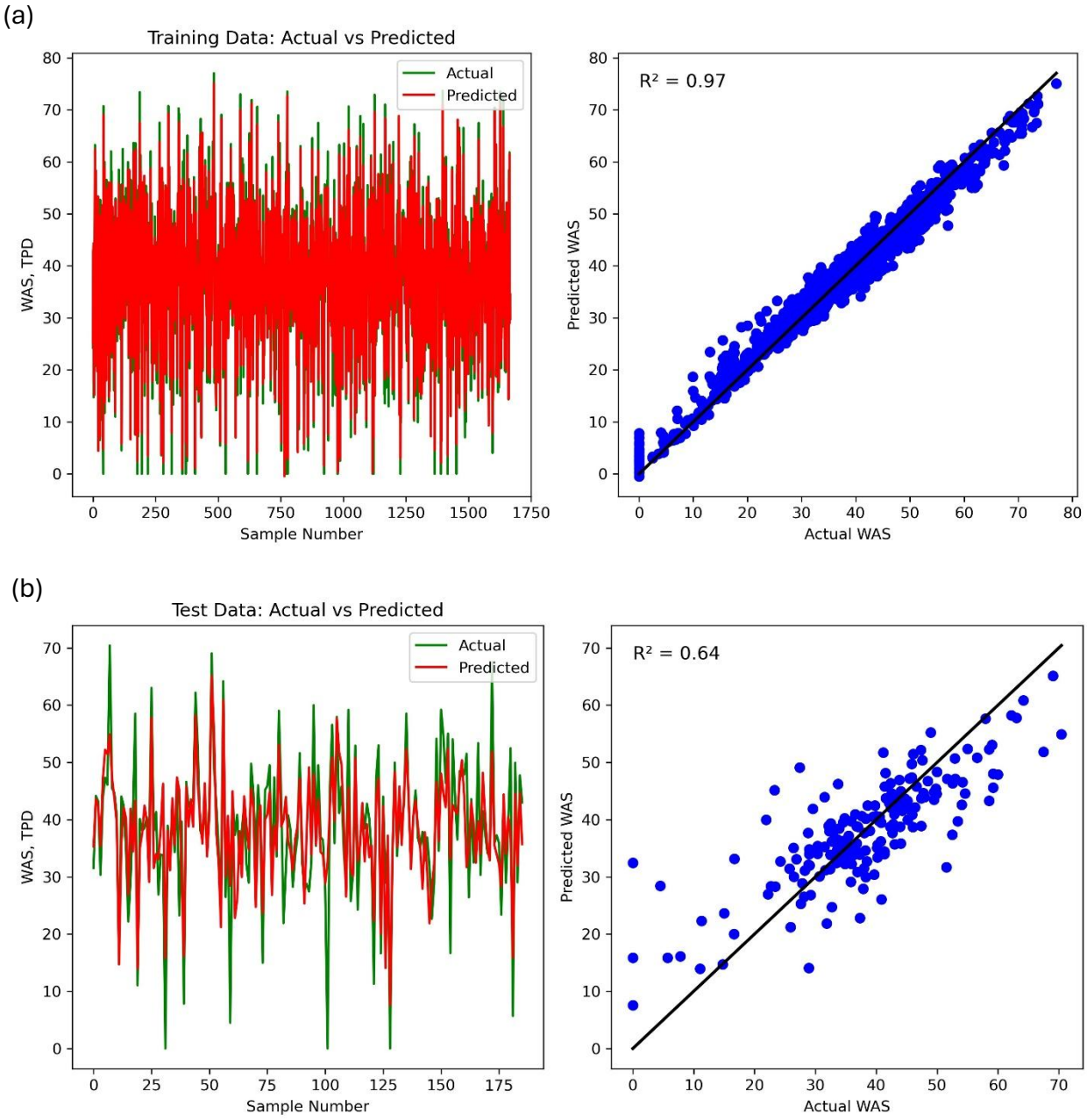

Figure 2.7. GBM model performance in predicting WAS. (a) training data (b) test data

Table 2.2. Metrices results of optimized model for Primary sludge (a-RF, b-GBM and c-GBTJ); MAE and RMSE is in TPD, MSE is in TPD^2 unit

(a)

FS method	No. of input	(90:10)								(80:20)								(70:30)							
		MSE (Training)	MSE (Test)	MAE (Training)	MAE (Test)	R^2 (Training)	R^2 (Test)	RMSE (Training)	RMSE (Test)	MSE (Training)	MSE (Test)	MAE (Training)	MAE (Test)	R^2 (Training)	R^2 (Test)	RMSE (Training)	RMSE (Test)	MSE (Training)	MSE (Test)	MAE (Training)	MAE (Test)	R^2 (Training)	R^2 (Test)	RMSE (Training)	RMSE (Test)
	56.00	9.75	53.23	2.24	5.22	0.94	0.65	3.12	7.30	10.18	60.84	2.31	5.87	0.94	0.63	3.19	8.05	10.10	73.18	2.27	6.12	0.94	0.59	3.18	8.55
SHAP	10.00	15.82	62.66	2.70	5.59	0.91	0.60	3.98	7.92	16.89	67.60	2.79	6.07	0.90	0.61	4.11	8.22	16.07	75.95	2.73	6.26	0.91	0.57	4.01	8.72
	5.00	23.88	75.77	3.45	6.30	0.86	0.51	4.89	8.70	44.41	82.28	4.80	6.74	0.74	0.53	6.64	9.07	37.29	88.63	4.59	6.98	0.78	0.50	6.11	9.41
	3.00	66.12	105.87	6.08	7.73	0.62	0.31	8.13	10.29	68.41	106.06	6.21	7.94	0.60	0.39	8.27	10.30	62.78	117.75	5.99	8.28	0.63	0.34	7.92	10.85
	2.00	71.68	111.55	6.43	8.02	0.59	0.27	8.47	10.56	68.65	114.44	6.40	8.20	0.60	0.34	8.29	10.70	68.16	126.93	6.28	8.58	0.60	0.29	8.26	11.27
LASSO	10.00	11.26	59.51	2.41	5.83	0.94	0.61	3.35	7.71	17.36	71.90	2.84	6.30	0.90	0.59	4.17	0.59	11.58	77.13	2.45	6.45	0.93	0.57	3.40	8.78
	5.00	46.42	86.88	5.29	7.14	0.73	0.43	6.81	9.33	47.78	98.52	5.25	7.42	0.72	0.43	6.91	9.93	47.27	108.95	5.01	7.79	0.72	0.39	6.88	10.44
	3.00	73.73	100.27	6.52	7.42	0.58	0.35	8.59	10.01	57.52	117.59	5.77	7.94	0.67	0.32	7.58	10.84	69.04	125.82	6.30	8.41	0.60	0.29	8.31	11.22
	2.00	87.40	104.56	7.16	7.45	0.50	0.32	9.35	10.23	84.86	124.18	7.07	8.29	0.51	0.29	9.21	11.14	83.61	133.21	7.08	8.71	0.51	0.25	9.14	11.54
MI	10.00	10.58	57.57	2.31	5.43	0.94	0.63	3.25	7.59	10.65	69.52	2.33	5.79	0.94	0.60	3.27	8.34	11.20	81.91	2.35	6.36	0.93	0.54	3.35	9.05
	5.00	26.08	77.81	3.58	6.49	0.85	0.49	5.11	8.82	26.64	92.60	3.61	7.02	0.85	0.47	5.16	9.62	23.81	101.01	3.42	7.35	0.86	0.43	4.88	10.05
	3.00	41.99	92.30	4.81	6.95	0.76	0.40	6.48	9.61	54.26	101.79	5.35	7.44	0.69	0.41	7.37	10.09	66.99	114.37	6.08	7.86	0.61	0.36	8.18	10.69
	2.00	41.50	93.68	4.79	7.02	0.76	0.39	6.44	9.68	57.40	102.68	5.50	7.46	0.67	0.41	7.58	10.13	64.83	114.28	5.98	7.89	0.62	0.36	8.05	10.69
PC	10.00	17.12	69.00	2.95	6.17	0.90	0.55	4.14	8.31	12.93	78.97	2.62	6.52	0.93	0.55	3.60	8.89	20.82	93.56	3.19	7.06	0.88	0.47	4.56	9.67
	5.00	55.33	93.00	5.52	6.95	0.68	0.40	7.44	9.64	56.73	91.41	4.95	7.06	0.73	0.47	6.84	9.56	53.86	104.53	5.44	7.63	0.68	0.41	7.34	10.22
	3.00	67.48	97.73	6.20	7.39	0.61	0.36	8.21	9.89	69.27	102.50	6.29	7.65	0.60	0.41	8.32	10.12	64.71	114.64	6.08	8.02	0.62	0.36	8.04	10.71
	2.00	73.68	1110.17	6.59	7.77	0.58	0.28	8.58	10.50	87.40	123.40	7.08	8.54	0.49	0.30	9.35	11.11	74.59	131.55	7.09	8.70	0.49	0.26	9.31	11.47
RF	10.00	10.21	54.66	2.25	5.23	0.94	0.64	3.19	7.39	10.67	61.82	2.29	5.66	0.94	0.64	3.23	7.86	18.07	74.59	2.88	6.18	0.89	0.58	4.25	8.64
	5.00	25.79	80.85	3.52	6.39	0.85	0.47	5.08	8.99	14.63	83.11	2.82	6.59	0.92	0.52	3.83	9.12	24.94	97.40	3.45	7.23	0.85	0.45	4.99	9.87
	3.00	64.12	102.37	6.08	7.45	0.63	0.33	8.01	10.12	60.96	100.64	6.01	7.54	0.65	0.42	7.81	10.03	61.48	115.35	5.97	8.00	0.64	0.35	7.84	10.74
	2.00	75.73	113.03	6.56	8.08	0.57	0.26	8.70	10.63	68.62	113.79	6.38	8.18	0.60	0.35	8.28	10.67	64.89	125.60	6.23	8.56	0.62	0.29	8.06	11.21

(b)

FS method	No. of input	(90:10)								(80:20)								(70:30)							
		MSE (Training)	MSE (Test)	MAE (Training)	MAE (Test)	R^2 (Training)	R^2 (Test)	RMSE (Training)	RMSE (Test)	MSE (Training)	MSE (Test)	MAE (Training)	MAE (Test)	R^2 (Training)	R^2 (Test)	RMSE (Training)	RMSE (Test)	MSE (Training)	MSE (Test)	MAE (Training)	MAE (Test)	R^2 (Training)	R^2 (Test)	RMSE (Training)	RMSE (Test)
	56.00	5.06	55.03	1.65	5.40	0.97	0.64	2.25	7.42	4.22	58.69	1.48	5.70	0.98	0.66	2.06	7.66	3.05	66.51	1.29	5.93	0.98	0.63	1.75	8.12
SHAP	10.00	35.30	72.43	4.41	6.17	0.80	0.53	5.94	8.51	27.37	76.61	3.88	6.57	0.84	0.56	5.23	8.75	23.07	82.00	3.53	6.73	0.86	0.54	4.80	9.06
	5.00	57.36	90.61	5.67	7.04	0.67	0.41	7.57	9.52	84.87	97.56	6.98	7.56	0.51	0.44	9.21	9.88	62.49	107.96	6.00	7.68	0.63	0.39	7.90	10.39
	3.00	75.88	99.08	6.65	7.27	0.57	0.36	8.71	9.95	89.05	106.66	7.22	7.78	0.48	0.39	9.43	0.39	86.81	114.76	7.13	8.03	0.49	0.36	9.31	10.71
	2.00	104.45	106.95	7.79	7.65	0.40	0.30	10.22	10.34	100.37	112.51	7.61	8.10	0.42	0.35	10.02	10.61	92.94	124.66	7.34	8.49	0.46	0.30	9.64	11.17
LASSO	10.00	11.55	63.53	2.52	6.11	0.93	0.59	3.40	7.97	4.40	73.45	1.55	6.49	0.97	0.58	2.09	8.57	12.26	72.96	2.57	6.50	0.93	0.59	3.50	8.54
	5.00	66.05	86.51	6.24	7.15	0.62	0.44	8.13	9.30	49.07	96.55	5.33	7.44	0.72	0.44	7.00	9.83	59.29	105.80	5.82	7.80	0.65	0.41	7.70	10.29
	3.00	106.10	101.66	7.94	7.49	0.39	0.34	10.30	10.08	89.70	120.54	7.26	8.15	0.48	0.31	9.47	10.98	99.02	123.98	7.67	8.35	0.42	0.30	9.95	11.13
	2.00	115.12	105.18	8.30	7.53	0.34	0.32	10.73	10.26	118.00	124.47	8.34	8.32	0.32	0.28	10.86	11.16	108.37	130.36	8.06	8.55	0.37	0.28	10.41	11.42
MI	10.00	4.94	57.10	1.62	5.56	0.97	0.63	2.22	7.56	4.67	61.06	1.55	5.76	0.97	0.65	2.16	7.81	2.97	73.53	1.21	6.34	0.98	0.59	1.72	8.57
	5.00	35.61	78.23	4.30	6.61	0.80	0.49	5.97	8.84	37.20	91.91	4.44	7.10	0.78	0.47	6.10	9.59	51.48	100.46	5.25	7.46	0.70	0.44	7.17	10.02
	3.00	54.84	106.31	5.41	7.63	0.69	0.31	7.41	10.31	62.23	107.73	5.76	7.57	0.64	0.38	7.89	10.38	59.84	110.92	5.66	7.72	0.65	0.38	7.74	10.53
	2.00	72.97	114.24	6.31	8.03	0.58	0.26	8.54	10.69	81.46	125.15	6.72	8.43	0.53	0.28	9.03	11.19	68.00	128.33	6.01	8.45	0.61	0.28	8.25	11.33
PC	10.00	33.58	80.83	4.25	6.73	0.81	0.47	5.80	8.99	37.59	86.24	4.54	6.89	0.78	0.50	6.13	9.29	32.74	94.11	4.27	7.25	0.81	0.47	5.72	9.70
	5.00	82.10	91.60	6.88	6.88	0.53	0.40	9.06	9.57	80.73	100.04	6.81	7.47	0.53	0.42	8.99	10.00	75.95	110.80	6.63	7.79	0.56	0.38	8.71	10.53
	3.00	93.67	95.91	7.39	7.24	0.46	0.38	9.68	9.79	93.25	106.03	7.41	7.75	0.46	0.39	9.66	10.30	87.71	117.77	7.19	8.15	0.49	0.34	9.37	10.85
	2.00	118.50	109.42	8.29	7.79	0.32	0.29	10.89	10.46	110.40	122.86	7.97	8.55	0.36	0.29	10.51	11.08	118.49	109.42	8.29	7.79	0.32	0.29	10.89	10.46
RF	10.00	16.93	57.35	2.99	5.22	0.90	0.63	4.11	7.57	17.52	65.29	3.05	5.77	0.90	0.62	4.19	8.08	5.11	74.39	1.65	6.27	0.97	0.58	2.26	8.62
	5.00	51.85	84.72	5.36	6.59	0.70	0.45	7.20	9.20	62.02	91.55	5.88	7.15	0.64	0.47	7.88	9.57	63.76	102.73	6.01	7.62	0.63	0.42	7.99	10.14
	3.00	92.54	97.72	7.33	7.21	0.47	0.36	9.62	9.89	89.04	106.60	7.22	7.78	0.48	0.39	9.44	10.32	86.81	114.95	7.13	8.03	0.49	0.35	9.32	10.72
	2.00	104.45	106.95	7.79	7.65	0.40	0.30	10.22	10.34	100.37	112.51	7.61	8.10	0.42	0.35	10.02	10.61	92.94	124.66	7.34	8.49	0.46	0.30	9.64	11.16

(c)

FS method	No. of input	(90:10)								(80:20)								(70:30)							
		MSE (Training)	MSE (Test)	MAE (Training)	MAE (Test)	R^2 (Training)	R^2 (Test)	RMSE (Training)	RMSE (Test)	MSE (Training)	MSE (Test)	MAE (Training)	MAE (Test)	R^2 (Training)	R^2 (Test)	RMSE (Training)	RMSE (Test)	MSE (Training)	MSE (Test)	MAE (Training)	MAE (Test)	R^2 (Training)	R^2 (Test)	RMSE (Training)	RMSE (Test)
	56.00	5.12	52.80	1.66	5.34	0.97	0.66	2.26	7.27	3.97	60.58	1.54	5.72	0.98	0.65	1.99	7.78	11.72	68.70	2.59	6.00	0.93	0.61	3.42	8.23
SHAP	10.00	13.57	67.91	2.72	5.74	0.92	0.56	3.68	8.24	14.97	68.71	2.82	6.16	0.91	0.60	3.87	6.29	9.43	78.79	2.27	6.28	0.94	0.55	3.07	8.88
	5.00	48.70	85.49	5.27	6.78	0.72	0.44	6.98	9.25	41.05	94.69	4.88	7.33	0.76	0.45	6.41	9.73	50.21	101.81	5.41	7.54	0.71	0.43	7.09	10.09
	3.00	90.50	102.89	7.26	7.41	0.48	0.33	9.51	10.14	96.02	107.96	7.45	7.86	0.44	0.38	9.80	10.39	82.42	118.85	6.97	8.16	0.52	0.33	9.08	10.90
	2.00	114.41	115.03	8.15	7.81	0.35	0.25	10.70	10.73	103.98	121.31	7.75	8.45	0.40	0.30	10.20	11.01	120.24	132.37	8.38	8.74	0.30	0.26	10.96	11.51
LASSO	10.00	14.91	65.98	2.98	6.19	0.91	0.57	3.86	8.12	14.91	72.75	2.87	6.50	0.91	0.58	3.86	8.53	13.41	76.26	2.74	6.65	0.92	0.57	3.66	8.73
	5.00	71.70	86.16	6.47	7.02	0.59	0.44	8.47	9.28	66.19	101.39	6.29	7.50	0.62	0.42	8.14	10.07	74.26	111.27	6.63	7.93	0.56	0.38	8.62	10.55
	3.00	90.79	98.32	7.35	7.34	0.48	0.36	9.53	9.92	88.16	117.60	7.19	8.06	0.49	0.32	9.39	10.84	96.50	121.69	7.57	8.28	0.43	0.32	9.82	11.03
	2.00	118.97	105.10	8.39	7.50	0.32	0.32	10.91	10.25	116.65	12.14	8.30	8.27	0.32	0.29	10.80	11.10	114.60	129.79	8.25	8.54	0.33	0.27	10.71	11.39
MI	10.00	14.91	61.04	2.84	5.71	0.91	0.60	3.86	7.81	13.46	73.45	2.76	6.14	0.92	0.58	3.67	8.57	11.43	81.43	2.51	6.50	0.93	0.54	3.80	9.02
	5.00	42.63	73.82	4.81	6.45	0.76	0.52	6.53	8.59	41.40	90.93	4.79	7.02	0.76	0.48	6.43	9.54	54.49	100.88	5.54	7.52	0.68	0.43	7.38	10.04
	3.00	53.18	96.06	5.34	7.15	0.70	0.38	7.29	9.80	51.07	103.19	5.23	7.41	0.70	0.41	7.15	10.16	78.37	114.20	6.56	7.92	0.54	0.36	8.85	10.69
	2.00	53.18	95.87	5.34	7.14	0.70	0.38	7.29	9.79	69.89	103.04	6.16	7.50	0.60	0.41	8.36	10.15	78.37	114.19	6.56	7.92	0.54	0.36	8.85	10.69
PC	10.00	39.28	83.19	4.65	6.89	0.78	0.46	6.27	9.12	37.52	86.80	4.56	6.95	0.78	0.50	6.13	9.32	30.72	98.56	4.13	7.37	0.82	0.45	5.54	9.93
	5.00	77.95	95.00	6.73	7.01	0.55	0.38	8.83	9.75	77.76	99.76	6.73	7.56	0.55	0.43	8.82	9.99	71.87	112.75	6.48	7.86	0.58	0.37	8.48	10.62
	3.00	101.83	98.62	7.73	7.25	0.42	0.36	10.09	9.93	88.16	106.44	7.18	7.78	0.49	0.39	9.39	10.32	96.25	120.18	7.54	8.19	0.44	0.32	9.81	10.96
	2.00	101.83	98.62	7.73	7.25	0.42	0.36	10.09	9.93	88.16	106.45	7.18	7.78	0.49	0.39	9.39	10.32	96.25	120.21	7.54	8.19	0.44	0.32	9.81	10.96
RF	10.00	14.11	59.32	2.78	5.21	0.92	0.61	3.76	7.70	11.84	66.34	2.61	5.77	0.93	0.62	3.44	8.15	13.78	78.43	2.76	6.45	0.92	0.56	3.71	8.86
	5.00	61.26	89.64	5.85	6.76	0.65	0.42	7.83	9.47	44.84	90.66	5.06	7.10	0.74	0.48	6.70	9.52	53.35	105.50	5.49	7.62	0.69	0.41	7.30	10.27
	3.00	90.50	102.89	7.26	7.41	0.48	0.33	9.51	10.14	96.02	107.96	7.45	7.86	0.44	0.38	9.80	10.39	82.42	118.85	6.97	8.16	0.52	0.33	908.00	10.90
	2.00	95.81	109.80	7.49	7.70	0.45	0.28	9.79	10.48	94.10	114.25	7.38	8.14	0.46	0.34	9.70	10.69	86.67	124.85	7.14	8.46	0.49	0.30	9.31	11.17

(a)

FS method	No. of input	(90:10)								(80:20)								(70:30)							
		MSE (Training)	MSE (Test)	MAE (Training)	MAE (Test)	R^2 (Training)	R^2 (Test)	RMSE (Training)	RMSE (Test)	MSE (Training)	MSE (Test)	MAE (Training)	MAE (Test)	R^2 (Training)	R^2 (Test)	RMSE (Training)	RMSE (Test)	MSE (Training)	MSE (Test)	MAE (Training)	MAE (Test)	R^2 (Training)	R^2 (Test)	RMSE (Training)	RMSE (Test)
SHAP	56.00	20.14	140.41	3.19	8.57	0.95	0.68	4.48	11.85	28.24	134.95	3.53	8.54	0.93	0.68	5.31	11.62	21.85	131.71	3.30	8.52	0.95	0.67	4.67	11.48
	10.00	21.13	138.29	3.26	8.28	0.95	0.69	4.60	11.76	21.72	146.36	3.32	8.61	0.95	0.65	4.67	12.10	21.55	140.89	3.29	8.78	0.95	0.64	4.64	11.87
	5.00	101.39	186.57	7.40	10.38	0.76	0.58	10.07	13.66	100.43	182.54	7.35	10.21	0.77	0.56	10.02	13.51	77.06	182.31	6.56	10.27	0.82	0.54	8.77	13.50
	3.00	123.10	209.33	7.90	10.54	0.71	0.53	11.09	14.47	114.94	199.96	7.40	10.50	0.73	0.52	10.72	14.14	106.32	200.63	7.60	10.59	0.76	0.49	10.31	14.16
	2.00	174.05	237.99	9.60	11.78	0.59	0.46	13.19	15.43	175.13	233.79	9.63	11.54	0.59	0.44	13.23	15.29	174.89	228.50	9.50	11.60	0.60	0.42	13.22	15.11
LASSO	10.00	21.10	157.84	3.44	9.51	0.95	0.64	4.59	12.56	33.85	150.85	4.03	9.40	0.92	0.64	5.82	12.28	24.20	141.11	3.60	9.18	0.94	0.64	4.92	11.88
	5.00	63.27	194.95	6.09	10.38	0.85	0.56	7.95	13.96	62.02	177.53	6.02	10.05	0.86	0.57	7.88	13.32	68.74	174.21	6.20	10.10	0.84	0.56	8.29	13.20
	3.00	95.99	230.50	7.67	11.54	0.77	0.48	9.80	15.18	113.84	216.99	8.31	11.49	0.73	0.48	10.67	14.73	79.54	223.22	7.03	11.61	0.82	0.43	8.91	14.94
	2.00	144.84	248.49	9.52	12.36	0.66	0.44	12.04	15.76	142.19	244.76	9.45	12.26	0.67	0.41	11.92	15.64	141.63	243.40	940.00	12.26	0.68	0.38	11.90	15.60
MI	10.00	20.79	145.87	3.27	8.52	0.95	0.67	4.56	12.08	21.59	143.66	3.29	8.65	0.95	0.65	4.65	11.99	2.22	138.22	3.36	8.55	0.95	0.65	4.71	11.76
	5.00	87.90	151.66	6.43	8.98	0.79	0.66	9.38	12.31	67.02	162.72	5.82	9.35	0.84	0.61	8.19	12.76	43.68	163.92	4.55	9.55	0.90	0.58	6.61	12.80
	3.00	153.44	229.50	9.09	11.49	0.64	0.48	12.39	15.15	156.66	228.11	9.20	11.44	0.63	0.45	12.52	15.10	155.60	216.39	9.03	11.44	0.65	0.45	12.47	14.71
	2.00	215.45	259.13	11.36	12.87	0.49	0.41	14.68	16.10	201.37	282.82	11.04	13.46	0.53	0.32	14.19	16.82	213.40	280.21	11.08	13.44	0.51	0.29	14.61	16.74
PC	10.00	23.58	174.00	3.57	9.88	0.94	0.61	4.85	13.19	38.61	169.03	4.26	9.76	0.91	0.59	6.21	13.00	116.51	182.91	8.01	10.48	0.73	0.54	10.79	13.52
	5.00	117.84	180.72	8.07	10.32	0.72	0.59	10.86	13.44	107.61	189.83	8.00	10.57	0.75	0.54	10.37	13.78	116.53	185.57	8.00	10.56	0.73	0.53	10.79	13.62
	3.00	139.64	216.57	8.67	10.95	0.67	0.51	11.82	14.72	135.91	216.87	8.57	10.94	0.68	0.48	11.66	14.73	139.72	202.53	8.55	10.91	0.68	0.49	11.82	14.23
	2.00	204.70	263.03	11.14	12.97	0.52	0.41	14.31	16.22	202.50	281.70	11.02	13.39	0.53	0.32	14.23	16.78	199.67	282.78	10.86	13.44	0.55	0.28	14.13	16.82
RF	10.00	23.62	173.44	3.56	9.93	0.94	0.61	4.86	13.17	38.84	173.72	4.31	9.94	0.91	0.58	6.23	13.18	39.79	163.34	4.35	9.65	0.91	0.59	6.31	12.78
	5.00	120.83	184.55	8.19	10.43	0.72	0.58	10.99	13.58	106.99	188.15	7.96	10.49	0.75	0.55	10.34	13.72	97.18	180.55	7.40	10.43	0.78	0.54	9.86	13.44
	3.00	155.73	225.27	9.18	11.46	0.63	0.49	12.48	15.01	158.09	225.62	9.22	11.42	0.63	0.46	12.57	15.02	159.31	215.47	9.15	11.39	0.64	0.45	12.62	14.68
	2.00	212.31	263.45	11.27	12.99	0.50	0.41	14.57	16.23	201.92	279.66	11.03	13.36	0.53	0.33	14.21	16.72	200.07	282.29	10.88	13.45	0.54	0.28	14.14	16.80

(b)

FS method	No. of input	(90:10)								(80:20)								(70:30)							
		MSE (Training)	MSE (Test)	MAE (Training)	MAE (Test)	R^2 (Training)	R^2 (Test)	RMSE (Training)	RMSE (Test)	MSE (Training)	MSE (Test)	MAE (Training)	MAE (Test)	R^2 (Training)	R^2 (Test)	RMSE (Training)	RMSE (Test)	MSE (Training)	MSE (Test)	MAE (Training)	MAE (Test)	R^2 (Training)	R^2 (Test)	RMSE (Training)	RMSE (Test)
SHAP	56.00	37.10	133.59	4.55	8.63	0.91	0.70	6.09	11.56	12.64	136.46	2.58	8.83	0.97	0.67	3.56	11.68	8.13	137.29	2.11	8.72	0.98	0.65	2.85	11.72
	10.00	37.66	149.54	4.56	8.85	0.91	0.66	6.13	12.22	63.11	142.23	5.90	8.90	0.85	0.66	7.94	11.93	68.28	139.95	6.23	8.93	0.84	0.65	8.26	11.83
	5.00	130.90	199.65	8.80	10.90	0.69	0.55	11.44	14.13	133.87	201.08	8.94	10.95	0.69	0.52	11.57	14.18	139.30	189.72	9.24	10.71	0.68	0.52	11.80	13.78
	3.00	214.84	221.20	11.12	11.51	0.49	0.50	14.66	14.87	203.80	207.78	10.77	11.09	0.52	0.50	14.28	14.41	224.45	200.66	11.34	11.01	0.49	0.49	14.98	14.17
	2.00	212.61	229.88	10.88	11.70	0.50	0.48	14.58	15.16	215.85	215.25	11.00	11.22	0.50	0.48	14.69	14.67	219.47	208.72	11.06	11.13	0.50	0.47	14.81	14.45
LASSO	10.00	37.45	164.81	4.63	9.56	0.91	0.63	6.12	12.84	59.85	152.70	5.93	9.65	0.86	0.63	7.74	12.36	58.70	146.18	5.84	9.46	0.87	0.63	7.67	12.09
	5.00	96.93	202.37	7.45	10.60	0.77	0.54	9.85	14.23	81.56	185.22	6.78	10.22	0.81	0.55	9.03	13.61	105.55	171.56	7.73	10.11	0.76	0.56	10.28	13.10
	3.00	159.61	228.71	9.89	11.48	0.63	0.49	12.63	15.12	169.16	218.28	10.21	11.57	0.61	0.48	13.01	14.77	141.35	217.17	9.22	11.62	0.68	0.45	11.89	14.74
	2.00	209.23	243.97	11.45	12.21	0.51	0.45	14.46	15.62	211.83	231.74	11.60	12.09	0.51	0.44	14.55	15.22	213.15	231.29	11.64	12.16	0.52	0.41	14.60	15.21
MI	10.00	37.96	132.98	4.54	8.59	0.91	0.70	6.16	11.53	68.38	138.18	6.19	8.88	0.84	0.67	8.27	11.75	29.25	139.86	3.96	8.95	0.93	0.65	5.41	11.83
	5.00	105.95	144.57	7.65	9.03	0.75	0.67	10.29	12.02	75.45	156.75	6.43	9.45	0.82	0.62	8.69	12.52	106.16	155.88	7.67	9.69	0.76	0.60	10.30	12.49
	3.00	214.84	221.20	11.12	11.51	0.49	0.50	14.66	14.87	203.80	207.78	10.77	11.09	0.52	0.50	14.28	14.41	224.45	200.66	11.35	11.01	0.49	0.49	14.98	14.17
	2.00	281.84	265.30	13.06	13.16	0.36	0.33	16.79	16.29	269.76	274.32	12.92	13.37	0.37	0.34	16.42	16.56	281.84	265.30	13.06	13.16	0.36	0.33	16.79	16.29
PC	10.00	46.22	179.96	5.05	10.07	0.89	0.59	6.80	13.41	88.52	173.24	7.16	9.97	0.79	0.58	9.41	13.16	88.92	174.46	7.15	10.05	0.80	0.56	9.43	13.21
	5.00	130.90	199.65	8.80	10.90	0.69	0.55	11.44	14.13	133.87	201.08	8.94	10.95	0.69	0.52	11.57	14.18	139.30	189.72	9.25	10.71	0.68	0.52	11.80	13.77
	3.00	189.62	200.77	10.41	10.74	0.55	0.55	13.77	14.17	172.30	198.31	9.95	10.74	0.60	0.52	13.13	14.08	191.18	195.74	10.52	10.97	0.57	0.50	13.83	13.99
	2.00	273.90	271.27	12.10	13.24	0.35	0.39	16.55	16.47	269.76	274.32	12.92	13.37	0.37	0.34	16.42	16.56	281.84	265.30	13.06	13.16	0.36	0.33	16.79	16.29
RF	10.00	46.22	179.86	5.05	10.06	0.89	0.60	6.80	13.41	88.52	173.35	7.16	9.98	0.79	0.58	9.41	13.17	102.30	170.71	7.71	9.98	0.77	0.57	10.11	13.07
	5.00	130.90	200.00	8.80	10.90	0.69	0.55	11.44	14.13	133.87	201.08	8.94	10.95	0.69	0.52	11.57	14.18	139.30	189.72	9.25	10.71	0.68	0.52	11.80	13.77
	3.00	214.84	221.20	11.12	11.51	0.49	0.50	14.66	14.87	203.80	207.78	10.77	11.09	0.52	0.50	14.28	14.42	224.45	200.66	11.35	11.01	0.49	0.49	14.98	14.17
	2.00	273.90	271.27	12.99	13.24	0.35	0.39	16.55	16.47	264.30	278.19	12.75	13.43	0.38	0.33	16.26	16.68	281.84	265.30	13.06	13.16	0.36	0.33	16.79	16.29

(c)

FS method	No. of input	(90:10)								(80:20)								(70:30)							
		MSE (Training)	MSE (Test)	MAE (Training)	MAE (Test)	R^2 (Training)	R^2 (Test)	RMSE (Training)	RMSE (Test)	MSE (Training)	MSE (Test)	MAE (Training)	MAE (Test)	R^2 (Training)	R^2 (Test)	RMSE (Training)	RMSE (Test)	MSE (Training)	MSE (Test)	MAE (Training)	MAE (Test)	R^2 (Training)	R^2 (Test)	RMSE (Training)	RMSE (Test)
SHAP	56.00	12.97	128.09	2.67	8.31	0.97	0.71	3.60	11.32	23.43	133.74	3.64	8.63	0.95	0.68	4.84	11.56	9.05	129.29	2.23	8.70	0.98	0.67	3.00	11.37
	10.00	35.98	141.32	4.46	8.55	0.92	0.68	6.00	11.89	65.67	139.36	6.26	8.80	0.85	0.67	8.10	11.81	52.06	138.24	5.53	8.94	0.88	0.65	7.22	11.76
	5.00	115.08	186.40	8.24	10.00	0.73	0.58	10.73	13.65	139.89	189.32	8.93	10.47	0.67	0.54	11.83	13.76	137.62	181.83	8.87	10.29	0.69	0.54	11.73	13.48
	3.00	196.48	215.71	10.63	11.42	0.54	0.51	10.02	14.69	190.51	214.84	10.63	11.24	0.56	0.48	13.80	14.66	206.90	202.46	11.09	11.03	0.53	0.49	14.38	14.23
	2.00	209.36	228.11	10.83	11.66	0.51	0.49	14.47	15.10	224.52	213.85	11.29	11.18	0.48	0.49	14.98	14.62	215.28	212.12	10.94	11.22	0.51	0.46	14.67	14.56
LASSO	10.00	61.91	149.25	6.11	9.38	0.85	0.66	7.86	12.22	53.80	148.89	5.72	9.50	0.87	0.64	7.33	12.21	41.04	145.25	4.92	9.27	0.91	0.63	6.41	12.05
	5.00	93.84	200.46	7.36	10.45	0.78	0.55	9.69	14.16	95.24	185.23	7.54	10.29	0.78	0.55	9.76	13.61	109.54	175.94	8.03	10.23	0.75	0.55	10.47	13.26
	3.00	132.12	238.00	9.12	11.70	0.69	0.46	11.49	15.43	158.77	218.12	9.96	11.53	0.63	0.48	12.60	14.77	157.15	217.40	9.85	11.60	0.64	0.45	12.54	14.74
	2.00	203.52	243.90	11.38	12.36	0.52	0.45	14.27	15.62	203.30	233.95	11.36	12.15	0.53	0.44	14.26	15.30	190.27	233.79	11.00	12.09	0.57	0.41	13.79	15.29
MI	10.00	56.37	137.38	5.75	8.79	0.87	0.69	7.51	11.72	44.90	136.93	5.02	8.88	0.90	0.67	6.70	11.70	21.46	148.06	3.43	9.16	0.95	0.62	4.63	12.17
	5.00	83.03	136.68	6.83	8.68	0.80	0.69	9.11	11.69	55.30	154.49	5.69	9.38	0.87	0.63	7.44	12.43	129.43	165.41	8.61	9.97	0.71	0.58	11.38	12.86
	3.00	196.48	215.71	10.63	11.42	0.54	0.51	14.02	14.69	190.51	214.71	10.63	11.23	0.56	0.48	13.80	14.65	206.90	202.34	11.09	11.02	0.53	0.49	14.38	14.22
	2.00	266.59	269.21	12.87	13.24	0.37	0.39	16.33	16.41	266.93	275.38	12.85	13.36	0.38	0.34	16.34	15.59	271.97	267.38	12.88	13.23	0.38	0.32	16.49	16.35
PC	10.00	103.28	174.29	7.79	9.93	0.76	0.61	10.16	13.20	66.85	174.01	6.20	10.11	0.84	0.58	8.18	13.19	82.48	167.10	7.06	9.76	0.81	0.58	9.08	12.93
	5.00	149.38	188.87	9.51	10.69	0.65	0.57	12.22	13.74	134.33	196.35	9.08	10.77	0.69	0.53	11.59	14.01	132.45	191.81	9.00	10.66	0.70	0.51	11.51	13.85
	3.00	183.26	197.63	10.32	10.69	0.57	0.56	13.54	14.06	196.14	200.99	10.82	10.98	0.54	0.52	14.00	14.18	178.89	192.65	10.30	10.82	0.59	0.51	13.37	13.88
	2.00	266.59	269.21	12.87	13.24	0.37	0.39	16.33	16.41	266.93	275.38	12.85	13.36	0.38	0.34	16.34	16.59	271.38	267.38	12.88	13.23	0.38	0.32	16.49	16.35
RF	10.00	103.28	174.17	7.79	9.92	0.76	0.61	10.16	13.20	66.85	173.96	6.20	10.12	0.84	0.58	8.18	13.19	71.07	168.79	6.49	9.78	0.84	0.57	8.43	13.00
	5.00	148.11	189.05	9.49	10.66	0.65	0.57	12.17	13.75	134.33	196.30	9.08	10.76	0.69	0.53	11.59	14.01	129.92	193.50	8.98	10.75	0.70	0.51	11.40	13.91
	3.00	196.48	215.53	10.63	11.42	0.54	0.51	14.02	14.68	190.51	214.74	10.63	11.23	0.56	0.48	13.80	14.65	206.90	202.73	11.09	11.04	0.53	0.49	14.38	14.24
	2.00	266.59	269.21	12.87	13.24	0.37	0.39	16.33	16.41	266.93	275.38	12.85	13.36	0.38	0.34	16.34	16.59	271.97	267.38	12.88	13.23	0.38	0.32	16.49	16.35

Table 2.2 (continued)

However, although increasing features can improve performance on the training set, it also decreases performance on the test dataset because of overfitting issues. Interestingly, 10 input variables achieved accuracy as high as of full dataset. According to paired t-test, for both GBM and GBT models, the results revealed that statistical metrics (MSE, MAE, R^2, and RMSE) differences of training and test data were not statistically significant for any combination of data split when input variables were selected by MI, PC, RF, and SHAP. It is understandable that GBT, being a variant of GBM, exhibited similar behavior. Because there was no statistically significant difference between training and test results, it suggests that GBM and GBT performed consistently across both training and test datasets. This consistency indicates that the model generalizes well to unseen data. The study also indicates that as a FS method, MI, PC, RF, and SHAP gave the best set of input variables to capture the complicated process. Although RF performed well as an FS method, it could not perform consistently across training and test datasets as a prediction model. In predicting WAS, RF, GBM, and GBT models were able to achieve R^2 values of 0.94, 0.98, and 0.98 for the training dataset and 0.65, 0.66, and 0.66 for the test dataset, respectively. The results revealed that metrics (MAE, R^2, and RMSE) differences for the training and test datasets were not statistically significant for GBM model (when variables were selected by MI and LASSO) and for GBT model (when variables were selected by PC and LASSO). The only difference of MSE was found for 70:30 data split in both models. With less training data, the model might be underfitting, which caused higher difference in MSE. However, the 90:10 and 80:20 splits might allow the model to capture more complex patterns in the data, resulting in lower MSE differences. Moreover, based on the findings of the study, the number of input variables and choice of FS method can significantly impact model

performance, as seen in the varying R^2 values across different FS methods and ML models. The difference in statistical significance between training and test data for different data splits can be attributed to the variability in the composition of the datasets used for training and testing the models. When the dataset is split into a larger portion for training (e.g., 90%) and a smaller portion for testing (e.g., 10%), some models might be overfitted to the training data. This means that the model captures the noise or random fluctuations in the training data, which may not generalize well to unseen data in the test set. As a result, the performance metrics (such as MSE, MAE, R^2, and RMSE) calculated on the test set might be significantly different from those on the training set, indicating poor generalization of the model. Nevertheless, when the dataset is split into a smaller portion for training (e.g., 70%) and a larger portion for testing (e.g., 30%), the model might not capture enough information from the training data to effectively generalize to unseen data. In this case, the model might underfit the data, leading to poor performance on both the training and test sets. The choice of input variables selected by different FS methods can also affect the model performance and its ability to generalize to unseen data. If certain input variables are more relevant or informative for predicting the target variable, then the model trained using those variables might perform better on both the training and test sets, leading to smaller differences in performance metrics between the two sets. GBM and GBT also performed superior in WWTP variable prediction in other studies (Shao et al., 2023; and Wei et al., 2023). Bagherzadeh et al. (2021) demonstrated that GBM effectively generalized dataset patterns, delivering superior performance on unseen data, highlighting its efficacy in predicting WWTP variables. Our results indicate that GBM model has significant advantage in training time, completing training more quickly than GBT (Figure 2.8). However,

while GBM can effectively generalize dataset patterns and perform well on unseen data, it requires careful cross-validation to avoid overfitting, as the model tends to overemphasize outliers while minimizing errors.

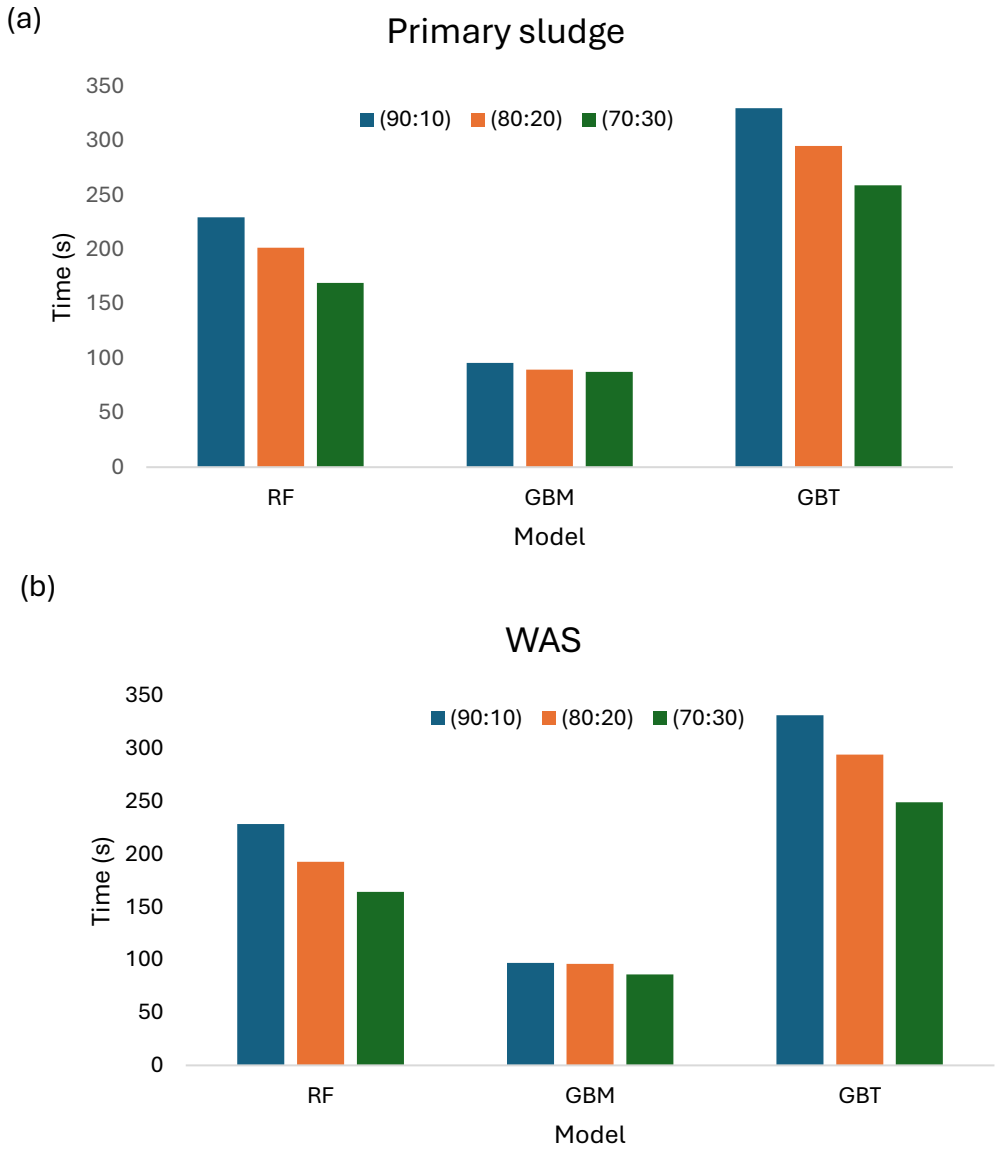

Figure 2.8. Computational time of ML models. (a) Primary sludge (b) WAS

2.4 Conclusions

This study presents advancement in the predictive modeling for sludge production in WWTP, utilizing ML alongside XAI techniques to enhance accuracy and interpretability. The findings demonstrate the efficacy of GBM and GBT in predicting sludge production. While GBM exhibited superior performance, the study emphasizes the significance of selecting appropriate input variables to capture intricate relationships and ensure robust predictions on unseen data. XAI techniques, SHAP and LIME, can identify key drivers of sludge production. This study serves as a practical demonstration of leveraging ML models and XAI to tackle real-world challenges in wastewater treatment, offering valuable insights for future endeavors in predictive modeling and process optimization. However, the study only included popular XAI methods SHAP and LIME. This limitation can be addressed by comparing their performance with other XAI methods in future research. In addition, XAI methods are in the early stage of development; the technique can be tested with different WWTP configurations to ensure their applicability and validity as demonstrated in our study. Future research could explore multi-site studies to enhance generalizability and incorporate real time sensor data for more dynamic and adaptive predictions in WWTP operations. Future studies could broaden the model's applicability and provide a more comprehensive understanding of sludge production in WWTPs by considering other types of sludge, such as chemical sludge.

2.5 References

1. Arnell, M., Ahlström, M., Wärff, C., Saagi, R., & Jeppsson, U. (2021). Plant-wide modelling and analysis of WWTP temperature dynamics for sustainable heat recovery from wastewater. Water Science and Technology, 84(4), 1023–1036.

2. Azimi, Y., Talaeian, M., Sarkheil, H., Hashemi, R., & Shirdam, R. (2022). Developing an evolving multi-layer perceptron network by genetic algorithm to predict full-scale municipal wastewater treatment plant effluent. Journal of Environmental Chemical Engineering, 10(5), 108398.

3. Bagherzadeh, F., Mehrani, M. J., Basirifard, M., & Roostaei, J. (2021). Comparative study on total nitrogen prediction in wastewater treatment plant and effect of various feature selection methods on machine learning algorithms performance. Journal of Water Process Engineering, 41, 102033.

4. Ching, P. M. L., Zou, X., Wu, D., So, R. H. Y., & Chen, G. H. (2022). Development of a wide-range soft sensor for predicting wastewater BOD5 using an eXtreme gradient boosting (XGBoost) machine. Environmental Research, 210, 112953.

5. Duarte, M. S., Martins, G., Oliveira, P., Fernandes, B., Ferreira, E. C., Alves, M. M., … Novais, P. (2023). A review of computational modeling in wastewater treatment processes. ACS ES&T Water, 4(3), 784–804.

6. Ekinci, E., Özbay, B., Omurca, S. İ., Sayın, F. E., & Özbay, İ. (2023). Application of machine learning algorithms and feature selection methods for better prediction of sludge production in a real advanced biological wastewater treatment plant. Journal of Environmental Management, 348, 119448.

7. El-Rawy, M., Abd-Ellah, M. K., Fathi, H., & Ahmed, A. K. A. (2021). Forecasting effluent and performance of wastewater treatment plant using different machine learning techniques. Journal of Water Process Engineering, 44, 102380.

8. Guo, H., Jeong, K., Lim, J., Jo, J., Kim, Y. M., Park, J. P., Kim, J. H., & Cho, K. H. (2015). Prediction of effluent concentration in a wastewater treatment plant using machine learning models. Journal of Environmental Sciences, 32, 90-101.

9. Hu, Y., Wei, R., Yu, K., Liu, Z., Zhou, Q., Zhang, M., … Qu, S. (2024). Exploring sludge yield patterns through interpretable machine learning models in China's municipal wastewater treatment plants. Resources, Conservation and Recycling, 204, 107467.

10. Jiang, M., Wang, J., Hu, L., & He, Z. (2023). Random forest clustering for discrete sequences. Pattern Recognition Letters, 174, 145-151.

Konstantinov, A. V., & Utkin, L. V. (2021). Interpretable machine learning with an ensemble of gradient boosting machines. Knowledge-Based Systems, 222, 106993.

11. Li, G., Ji, J., Ni, J., Wang, S., Guo, Y., Hu, Y., … Li, Y. Y. (2022). Application of deep learning for predicting the treatment performance of real municipal wastewater based on one-year operation of two anaerobic membrane bioreactors. Science of the Total Environment, 813, 151920.

12. Lundberg, S. M., Erion, G. G., & Lee, S. I. (2018). Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888.10.48550/arXiv.1802.03888.

13. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765-4774.

14. Ly, Q. V., Truong, V. H., Ji, B., Nguyen, X. C., Cho, K. H., Ngo, H. H., & Zhang, Z. (2022). Exploring potential machine learning application based on big data for prediction of wastewater quality from different full-scale wastewater treatment plants. Science of the Total Environment, 832, 154930.

15. Nourani, V., Elkiran, G., & Abba, S. I. (2018). Wastewater treatment plant performance analysis using artificial intelligence-An ensemble approach. Water Science and Technology, 78(10), 2064–2076.

16. Park, J., Lee, W. H., Kim, K. T., Park, C. Y., Lee, S., & Heo, T. Y. (2022). Interpretation of ensemble learning to predict water quality using explainable artificial intelligence. Science of the Total Environment, 832, 155070.

17. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). "Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135-1144).

18. Safder, U., Kim, J., Pak, G., Rhee, G., & You, K. (2022). Investigating machine learning applications for effective real-time water quality parameter monitoring in full-scale wastewater treatment plants. Water, 14(19), 3147.

19. Shao, S., Fu, D., Yang, T., Mu, H., Gao, Q., & Zhang, Y. (2023). Analysis of machine learning models for wastewater treatment plant sludge output prediction. Sustainability, 15(18), 13380.

20. Sun, J., Xu, Y., Nairat, S., Zhou, J., & He, Z. (2023). Prediction of biogas production in anaerobic digestion of a full-scale wastewater treatment plant using ensembled machine learning models. Water Environment Research, 95(6), e10893.

21. Sun, Z., Wang, G., Li, P., Wang, H., Zhang, M., & Liang, X. (2024). An improved random forest based on the classification accuracy and correlation measurement of decision trees. Expert Systems with Applications, 237, 121549.

22. Szomolányi, O., & Clement, A. (2023). Use of random forest for assessing the effect of water quality parameters on the biological status of surface waters. GEM-International Journal on Geomathematics, 14(1), 20.

23. Tung, T. M., & Yaseen, Z. M. (2020). A survey on river water quality modelling using artificial intelligence models: 2000–2020. Journal of Hydrology, 585, 124670.

24. Tyralis, H., Papacharalampous, G., & Langousis, A. (2019). A brief review of random forests for water scientists and practitioners and their recent history in water resources. Water, 11(5), 910.

25. Wang, D., Thunéll, S., Lindberg, U., Jiang, L., Trygg, J., Tysklind, M., & Souihi, N. (2021). A machine learning framework to improve effluent quality control in wastewater treatment plants. Science of the Total Environment, 784, 147138.

26. Wang, H. C., Wang, Y. Q., Wang, X., Yin, W. X., Yu, T. C., Xue, C. H., & Wang, A. J. (2024). Multimodal machine learning guides low carbon aeration strategies in urban wastewater treatment. Engineering, 36, 51–62.

27. Wei, X., Yu, J., Tian, Y., Ben, Y., Cai, Z., & Zheng, C. (2023). Comparative performance of three machine learning models in predicting influent flow rates and nutrient loads at wastewater treatment plants. ACS ES&T Water, 4(3), 1024-1035.

28. Wongburi, P., & Park, J. K. (2022). Prediction of sludge volume index in a wastewater treatment plant using recurrent neural network. Sustainability, 14(10), 6276.

29. Xie, Y., Chen, Y., Lian, Q., Yin, H., Peng, J., Sheng, M., & Wang, Y. (2022). Enhancing real-time prediction of effluent water quality of wastewater treatment plant based on improved feedforward neural network coupled with optimization algorithm. Water, 14(7), 1053.

30. Xu, Y., Wang, Z., Nairat, S., Zhou, J., & He, Z. (2024). Artificial intelligence-assisted prediction of effluent phosphorus in a fullscale wastewater treatment plant with missing phosphorus input and removal data. ACS ES&T Water, 4(3), 880–889.

31. Xu, Y., Zeng, X., Bernard, S., & He, Z. (2022). Data-driven prediction of neutralizer pH and valve position towards precise control of chemical dosage in a wastewater treatment plant. Journal of Cleaner Production, 348, 131360.

32. Yadav, P., Chandra, M., Fatima, N., Sarwar, S., Chaudhary, A., Saurabh, K., & Yadav, B. S. (2023). Predicting influent and effluent quality parameters for a UASB-based wastewater treatment plant in Asia covering data variations during COVID-19: A machine learning approach. Water, 15(4), 710.

33. Yu, J., Tian, Y., Jing, H., Sun, T., Wang, X., Andrews, C. B., & Zheng, C. (2023). Predicting regional wastewater treatment plant discharges using machine learning and population migration big data. ACS ES&T Water, 3(5), 1314–1328.

34. Zeinolabedini, M., & Najafzadeh, M. (2019). Comparative study of different wavelet-based neural network models to predict sewage sludge quantity in wastewater treatment plant. Environmental Monitoring and Assessment, 191(3), 163.

35. Zhang, S., Wang, H., & Keller, A. A. (2021). Novel machine learning-based energy consumption model of wastewater treatment plants. ACS ES&T Water, 1(12), 2531–2540.

36. Zhang, X., & Liu, C. A. (2023). Model averaging prediction by K-fold cross-validation. Journal of Econometrics, 235(1), 280-301.

37. Zhu, J. J., Borzooei, S., Sun, J., & Ren, Z. J. (2022). Deep learning optimization for soft sensing of hard-to-measure wastewater key variables. ACS ES&T Engineering, 2(7), 1341–1355.

CHAPTER 3: COMPARATIVE ANALYSIS OF MACHINE LEARNING MODELS AND EXPLAINABLE ARTIFICIAL INTELLIGENCE FOR PREDICTING WASTEWATER TREATMENT PLANT VARIABLES

3.1 Introduction

Wastewater treatment plants (WWTPs) play an essential role in safeguarding the aquatic environment by processing municipal and industrial sewage. Increasing amount of urban wastewater and demands for clean water present substantial challenges to WWTP operators in meeting regulatory effluent standards and reducing operating costs (Torregrossa et al., 2016; Abba et al., 2017; Bernardelli et al., 2020; Zhang et al., 2021). Moreover, the complexity of the treatment process demands a high level of precision to achieve the desired standard limits of various variables. To enhance effluent quality and comply with regulatory standards at WWTP while minimizing operation and maintenance cost, the implementation of advanced technologies is crucial. There is a potential for WWTPs to improve decision-making process and to optimize resource allocation by utilizing machine learning (ML), a subfield of artificial intelligence (AI) that can ultimately assist in achieving sustainable treatment system.

The application of ML in predicting WWTP variables has been effective (Zhang et al., 2021; Guo et al., 2015; Wang et al., 2021; El-Rawy et al., 2021; Li et al., 2022; Aghdam et al., 2023; Shyu et al., 2023; Wei et al., 2023; Xu et al., 2023; Yu et al., 2023; Alsulaili et al., 2021; Fan et al., 2018). ML models were also used to regulate WWTP operation that resulted in a notable amount of energy savings (Adibimanesh et al., 2023). According to studies Keerio et al., 2024; Solangi et al., 2024) ML can process substantial datasets with impressive precision.

As WWTPs are complex and comprise several concurrent nonlinear mechanisms, researchers investigated a wide range of variables, such as water quality, water quantity, and

meteorological data, in predicting WWTP variables using various ML models (Xu et al., 2023). Biochemical Oxygen Demand (BOD) and Total Suspended Solids (TSS) are among the most influential variables in WWTP. They were commonly investigated together because they share many similarities, including their hardness to measure, lack of information that may be obtained, the potential for complex model nonlinearity, and importance in prediction models (Zhu et al., 2022). Other common pollutants in wastewater are ammonia (NH_3) and phosphorus (P), both need to be reduced to the required level before being released into the environment (Bagherzadeh et al., 2021). A thorough understanding of nutrient characteristics is essential for the optimization of treatment operations (Wu et al., 2022; Keerio et al., 2020). Therefore, accurate effluent variable (BOD, NH_3, P, and TSS) prediction through ML can facilitate efficient adjustment of operational parameters such as aeration rates or chemical dosages to effectively meet effluent quality standards.

ML-based approaches are specifically being employed for the monitoring and design of complex non-linear issues at WWTPs (Singh et al., 2023). Traditional methods of variable measurements in WWTP involve time-consuming laboratory analysis. Advancements in sensor technologies and online monitoring systems have introduced real-time and alternative approaches. The difficulty of measuring BOD online and the length of time required for laboratory measurements highlight the significance of developing predictive models that can eliminate the requirement for measurements performed by humans. ML methods can rely on the connection created between the input and output datasets by extracting correlations between variables from historical data. Previous studies on various ML models to predict WWTP variables have a large variability in results, with R^2 for BOD ranging from 0.48 to 0.99,

TSS ranging from 0.63 to 0.98, (NH_3) ranging from 0.32-0.84, and P ranging from 0.28-0.93 (Abba et al., 2017; Wang et al., 2021; Wei et al., 2023; Alsulaili et al., 2021; Zhao et al., 2012; Bagheri et al., 2016; Sharghi et al., 2019; Khatri et al., 2019; Al-Ghazawi et al., 2021; Elmaadawy et al., 2021; Nourani et al., 2021; Ly et al., 2022; Dantas et al., 2023). Moreover, relying solely on ML models without an understanding of the contexts of the predictions is not ideal. Recent trend towards the practice of ML models in variable prediction requires explainability in addition to prediction accuracy. This is especially important in WWTPs where operators need to understand the reasons behind model predictions to increase their confidence in real-world application. Questions on rationale behind ML predictions, the basis for trust in these predictions, and methods for error correction are some of the concerns especially relevant in WWTP, where the reliability of ML practices is critical. While many studies have focused on predicting variables in WWTP using ML, research on implementing explainable artificial intelligence (XAI) is still developing. Some recent studies integrated XAI into interpreting ML output (Xu et al., 2023; Mahanna et al., 2024; Park et al., 2022). However, investigation of XAI methods with various ML models is lacking. Therefore, it is a novel attempt to investigate multiple XAI approaches to enhance the interpretability of ML models applications in WWTP.

This study applied XAI methods to improve the interpretability of ML models in predicting influential variables of a WWTP. Various feature selections and XAI methods were employed to identify the importance of input variables in ML models performance. A broad range of WWTP variables, encompassing water quality, water quantity, and electrical data were collected for the study. Several standalone ML models i.e. artificial neural network (ANN), gradient boosting machine (GBM), random forest (RF), eXtreme gradient boosting (XGBoost),

and hybrid model RF-GBM performance were tested and compared with historical datasets in predicting influent and effluent BOD, NH_3, P, and TSS. This study provides a better understanding of ML model performance in predicting WWTP variables with the help of XAI, which aids in making informed decisions to optimize treatment plant performance.

3.2 Materials and Methods

3.2.1 Data collection

The data were collected from a WWTP in Milwaukee, Wisconsin, USA that treats wastewater from industrial, municipal, and domestic sources. Water quality, water quantity, and electrical data (daily and hourly) were collected from 1st January 2019 to 31st December 2023. After data processing, the following variables were considered in the study: Influent BOD (BOD_i), Effluent BOD (BOD_e), Influent Flow ($Flow_i$), Effluent Flow ($Flow_e$), Influent Ammonia $(NH_3)_i$, Effluent Ammonia $(NH_3)_e$, Influent TSS (TSS_i), Effluent TSS (TSS_e), Influent Phosphorus (P_i), Effluent Phosphorus (P_e), TSS_e Removed, BOD_e Removed, Primary Sludge, Iron Dose, Detention Time, Aeration (Aer) Basin Temp, DO Set Pt, Sludge Volume Index (SVI), Mean Cell Residence Time (MCRT), Waste Activated Sludge (WAS), WAS Flow, pH_e, $Temp_e$, Total Residual Chlorine (TRC), Gravity Belt Thickening (GBT) Polymer Used, Fecal Coliforms, E.coli, Total Electricity (Elec) Generated, and Total Blower Elec Used. Time series of some significant variables can be found in Figure 3.1.

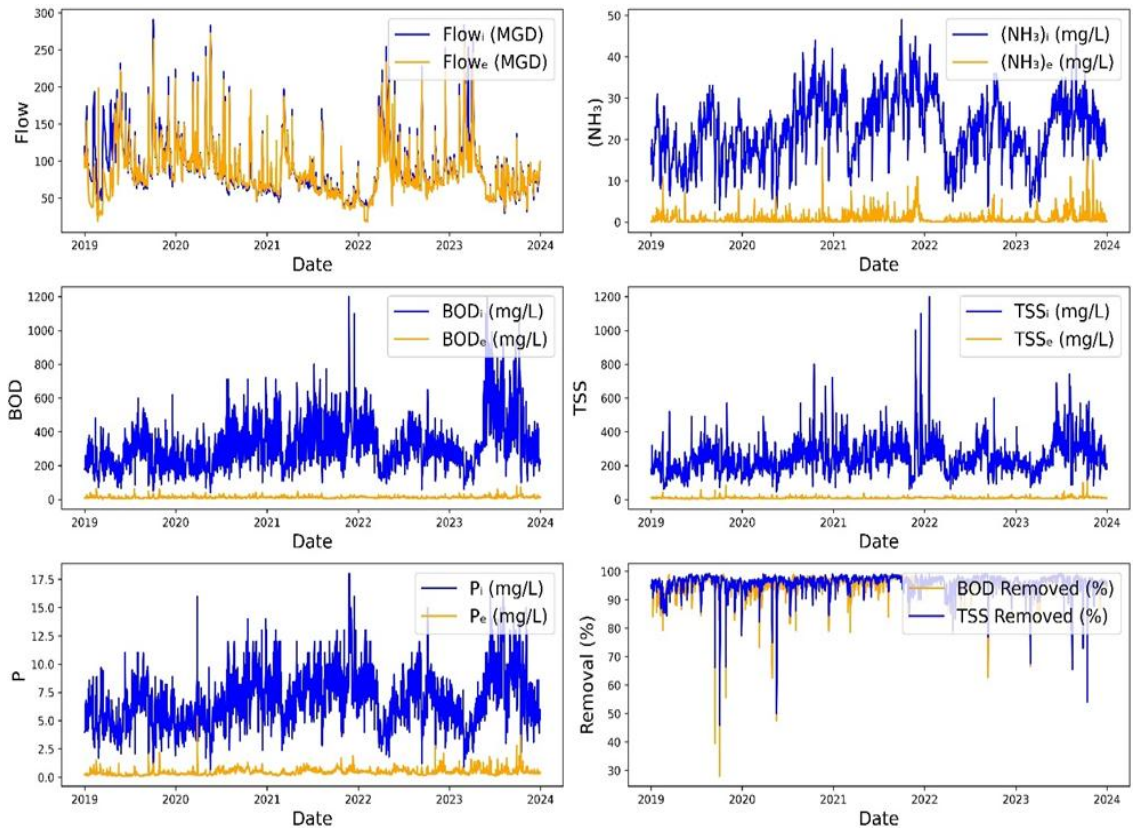

Figure 3.1. Time series of variables (top left: Flow; top right: (NH_3); middle left: BOD; middle right: TSS; bottom left: P; bottom right: BOD and TSS removed (%)).

3.2.2 Data Pre-Processing

Typically, sensor-collected data contains anomalies related to the recording process. During examination of the dataset for missing or inaccurate data, several anomalies were identified through human observation and subsequently replaced with average values. Additionally, any missing values in the dataset were filled in using the average value of the respective variable. We converted hourly variables to daily variables. The variables $Flow_i$ and $Flow_e$ exhibited a high correlation (0.9). To minimize multicollinearity, only $Flow_i$ was included in the study. Consequently, 28 out of the 29 collected variables consisting of 51128 data entries were considered for analysis. Statistical properties of data are presented in Table 3.1. When

eliminating redundant or irrelevant features that do not significantly affect the prediction, lowers noise, and enhances model performance (Sharghi et al., 2019), it is crucial to consider the context in which the model is used. Variables such as DO set points controlled by blowers may have a more indirect impact on the prediction accuracy, as they influence the performance of the overall treatment process rather than directly correlating with target variables. Moreover, outliers were not identified or removed in the dataset to understand the whole picture of the analysis as suggested by other studies (Hu et al., 2024). Therefore, in the study, full dataset of WWTP was considered that includes the most common input variables found in relevant papers to run the ML models (Shao et al., 2023).

Table 3.1. Data sets statistical properties.

Variables	Units	Min	Max	Mean	Std
$Flow_i$	(MGD)	29.58	290.95	89.18	39.32
$(NH_3)_i$	(mg/L)	2.90	49.00	22.36	7.73
$(NH_3)_e$	(mg/L)	0.02	18.00	0.95	1.56
BOD_i	(mg/L)	40.00	1200.00	331.86	155.69
BOD_e	(mg/L)	2.00	80.00	12.96	6.69
TSS_i	(mg/L)	46.00	1200.00	252.72	104.33
TSS_e	(mg/L)	1.90	110.00	9.23	6.21
P_i	(mg/L)	0.67	18.00	6.92	2.59
P_e	(mg/L)	0.09	5.40	0.47	0.32
TSS Removed	(%)	45.83	99.21	95.70	3.87
BOD Removed	(%)	28.00	99.14	95.16	4.36
Primary sludge	(TPD)	1.04	192.20	54.25	20.69
Iron Dose	(mg/L)	0.00	29.81	11.06	4.65
Detention Time	(min)	25.07	265.87	94.62	34.67
Aer Basin Temp	(F)	45.30	83.50	59.61	5.67
DO Set Pt	(mg/L)	3.00	5.00	3.63	0.36
SVI	(mL/g)	39.25	332.50	114.12	39.58
MCRT	(Days)	4.05	28.14	10.27	2.59
WAS	(TPD)	0.00	77.04	37.79	13.13
WAS Flow	(MGD)	0.00	2.88	1.61	0.45
pH_e	-	6.75	7.71	7.17	0.09
$Temp_e$	(F)	48.33	228.27	158.57	83.24
TRC	(mg/L)	0.00	0.04	0.01	0.01

GBT Polymer Used	(lbs/day)	0.00	86402.40	5107.26	5687.10
Fecal Coliforms	(CFU/100ml)	2.00	30000.00	306.74	1818.53
E coli.	(MPN/100ml)	1.00	24000.00	1017.97	8734.47
Total Elec Generated	(MW)	0.00	5.09	3.34	0.82
Total Blower Elec Used	(KW)	1371.38	3406.56	2859.10	331.63

3.2.3 Feature Selection

Several feature selection (FS) methods were employed to identify the most significant variables for predicting target variables, including analysis of variance (ANOVA), least absolute shrinkage and selection operator (LASSO), mutual information (MI), random forest (RF), and Pearson correlation (PC) (Xu et al., 2023; Bagherzadeh et al., 2021). ANOVA F-values are non-negative and can theoretically range from 0 to infinity. LASSO scores can be negative or positive, whereas PC scores range from -1 to 1. The MI scores range from 0, indicating no shared information, to positive values. RF generates a feature importance score from 0 to 1, where 0 means the feature was not used in the prediction, and 1 means the feature perfectly predicted the output. Traditional FS methods were chosen to compare their derived results with XAI method outputs.

3.2.4 SHapley Additive exPlanations

SHapley Additive exPlanations (SHAP) analysis is a recently developed XAI method based on game theory that interprets the behavior of ML models (Xu et al., 2023; Shao et al., 2023; Shafighfard et al., 2024; Lundberg et al., 2018). It explains the models' predictions by showcasing the relative influence of input variables on model performance (Mahanna et al., 2024). Using Shapley values from game theory, each feature is attributed values, as described by (Ludberg et al., 2017; Lundberg et al., 2018; Li et al., 2024) as follows:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|s|!\,(n - |s| - 1)!}{n!} [f_x(sU\{i\}) - f_x(s)] \quad (1)$$

Where ϕi is the SHAP value of ith input feature, n is the number of all input features, s is the subset of feature subsets, |s| is the feature subsets element number, fx(sU{i}) is trained with that feature present, and fx(s) is trained with feature withheld.

SHAP values at higher positions signify a greater importance of input variables on the models' performance. A positive weight indicates that increasing the feature's value typically boosts the models' prediction, whereas a negative weight implies that increasing the feature's value tends to reduce the model's prediction. SHAP summary plots are being used in WWTP to interpret models' output (Xu et al., 2023). In this study, we chose the commonly used function, SHAP summary plot, to investigate how the top features in a dataset impact the models' output.

3.2.5 Local Interpretable Model-Agnostic Explanation

Local Interpretable Model-Agnostic Explanation (LIME) is an XAI tool that interprets black-box ML models by using a local, interpretable model to clarify each prediction (Park et al., 2022). LIME is obtained by following equation:

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \mathcal{L}(f, g, \pi_x) + \Omega(g) \quad (2)$$

Where, $\mathcal{L}$ indicates fidelity function, G indicates explanation families, and Ω indicates complexity measure. The explanation model for instance x is the model g, π_x indicates proximity measure and f indicates original model.

LIME identifies the top features contributing most to the model's predictions, associating each feature with a weight that indicates its impact on the prediction. Features with positive weights have a positive effect on the prediction, while those with negative weights have a negative effect. The magnitude of the weight reflects the strength of the feature's influence on the prediction. Features are ranked by their importance, with the most influential ones listed first. A detailed explanation of LIME is provided by Riberio et al., 2016.

3.2.6 ML Models

To predict BOD_e, $(NH_3)_e$, P_e, and TSS_e, several ML models, i.e., ANN, GBM, RF, RF-GBM, and XGBoost were applied. These models were chosen because of their widespread application in water quality variable prediction. ANN consists of layered networks of interconnected nodes, with multiple hidden layers that allow the identification of intricate relationships and patterns in the data (Ye et al., 2020; Matheri et al., 2021). However, it requires substantial data and careful hyperparameter tuning. Different configurations of hidden layers, activation functions, and optimization strategies were explored in the study to train ANN model. A boosting approach called GBM combines several weak prediction models, typically decision trees, to produce a powerful predictive model (Konstantinov and Utkin 2021). To fix the errors created by the previous trees, GBM iteratively adds new models. Different values for learning rate, number of trees, tree depth, min sample leaf, and minimum sample split were used to identify the best combination that optimizes model performance on the training data. An ensemble learning technique called RF uses several decision trees to produce predictions (Tyralis et al., 2019; Nafsin et al., 2023; Jiang et al., 2023; Szomolányi et al., 2023; Sun et al., 2024). In this study, various values for number of trees, tree depth, min sample leaf, and minimum sample

split were used to identify the best combination that optimizes model performance on the training data. RF-GBM combines the principles of RF and GBM. This hybrid model combines the advantages of RF and GBM to improve prediction performance. A newly developed version of the gradient boosting decision tree algorithm called XGBoost has the potential to reduce overfitting and increase robustness (Shao et al., 2023). In XGBoost, several hyperparameters were also tuned to find that optimal configuration. In all ML models', the GridSearchCV is employed to identify the best combination of hyperparameters by testing multiple combinations using cross-validation.

3.2.7 Model Training and Evaluation

The dataset was divided into training and testing sets to ensure that the models were trained on a representative subset and evaluated on unseen data, providing a reliable measure of their generalization capability (Yadav et al., 2023). Two commonly recommended splits of training and test set ratios (90:10 and 80:20) were used as suggested by other relevant studies (Xu et al., 2023; Hu et al., 2024; Sun et al., 2024). The testing set acts as an independent dataset to assess the performance of the models, while the training set was utilized to train multiple ML models. Validation is a crucial step of the model development process to ensure that the developed model is accurate enough for the intended use (Xie et al., 2022; Sargent 2010; Tsioptsias et al., 2016). For validation purposes, splitting the data guarantees that the models are trained on a representative subset of the data and evaluated on unseen data, giving a trustworthy assessment of their generalization ability (Yadav et al., 2023). A 5-fold cross-validation was implemented to confirm the model's accuracy for its intended application by dividing the dataset into five equal parts (Zhang and Liu 2023). In each iteration, a different fold

was used as the test set, while the remaining folds constituted the training set. The model's performance is dependent on the hyperparameters used during training. To identify the best hyperparameter configuration, a grid search method was employed to find the optimal set that delivered the best performance.

To evaluate the regression model's performance, several model metrics can be used depending on the specific tasks, data characteristics, and circumstances (Kazemi et al., 2024; Bagherzadeh et al., 2023; Shafighfard 2022). In this regression study, three widely used assessment metrics-R-squared (R^2), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE)-were used to evaluate the performances of the ML models. MAE measures the average magnitude of the errors between predicted and actual values (eq 3). R^2 quantifies the percentage of variance that is explained by the models (eq 4), whereas RMSE denotes the average size of the residuals (eq 5). These metrics reveal information about the produced ML models' precision, goodness-of-fit, and accuracy. Higher values of R^2 and lower values of the error measures indicate better prediction performance and accuracy (Safder et al., 2022).

$$MAE = \frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{n} \tag{3}$$

$$R^2 = 1 - \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}} \tag{4}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2} \tag{5}$$

3.3 Results

3.3.1 ML Model Performance

The performance of ML models, including ANN, GBM, RF, XGBoost, and a hybrid RF-GBM, was evaluated using 90:10 and 80:20 train-test splits. The comparison between training and test performance helps to evaluate the models' generalization ability. An exceptionally high training performance relative to the test performance could be a sign of overfitting. Table 3.2 shows the model performance metrics for BOD prediction.

Table 3.2. Model performance metrices for 90:10 and 80:20 train-test splits.

			(90:10)			(80:20)		
Target Variable	Model	Set	MAE	R^2	RMSE	MAE	R^2	RMSE
BOD_e	ANN	Training	0.09	1.00	0.14	0.09	1.00	0.14
		Test	0.60	0.96	1.25	0.60	0.96	1.25
	GBM	Training	0.28	1.00	0.37	0.27	1.00	0.37
		Test	0.72	0.95	1.35	0.68	0.94	1.35
	RF	Training	0.45	0.99	0.77	0.48	0.99	0.79
		Test	1.27	0.86	2.36	1.18	0.86	2.07
	XGBoost	Training	0.33	1.00	0.45	0.29	1.00	0.39
		Test	0.74	0.95	1.36	0.69	0.94	1.36
	RF-GBM	Training	0.21	1.00	0.28	0.30	1.00	0.38
		Test	0.68	0.96	1.32	0.69	0.95	1.28
$(NH_3)_e$	ANN	Training	0.16	0.98	0.23	0.11	0.99	0.16
		Test	0.53	0.33	1.06	0.46	0.35	0.89
	GBM	Training	0.16	0.98	0.25	0.24	0.95	0.36
		Test	0.48	0.60	0.82	0.47	0.45	0.82
	RF	Training	0.20	0.94	0.40	0.21	0.94	0.40
		Test	0.51	0.52	0.90	0.48	0.48	0.79
	XGBoost	Training	0.05	1.00	0.07	0.03	1.00	0.04
		Test	0.47	0.57	0.85	0.43	0.55	0.74
	RF-GBM	Training	0.18	0.97	0.26	0.22	0.96	0.31
		Test	0.52	0.50	0.91	0.46	0.42	0.84
P_e	ANN	Training	0.11	0.73	0.17	0.10	0.74	0.17
		Test	0.14	0.42	0.18	0.13	0.44	0.19
	GBM	Training	0.01	1.00	0.01	0.03	0.99	0.04
		Test	0.10	0.65	0.14	0.11	0.55	0.17
	RF	Training	0.04	0.93	0.08	0.05	0.92	0.10
		Test	0.11	0.61	0.15	0.11	0.58	0.16

	XGBoost	Training	0.02	0.99	0.03	0.01	0.99	0.02
		Test	0.10	0.64	0.14	0.10	0.58	0.16
	RF-GBM	Training	0.01	1.00	0.01	0.01	1.00	0.01
		Test	0.10	0.65	0.14	0.10	0.60	0.16
TSS_e	ANN	Training	0.21	1.00	0.28	0.15	1.00	0.20
		Test	0.52	0.95	0.91	0.56	0.94	0.96
	GBM	Training	0.17	1.00	0.23	0.03	1.00	0.04
		Test	0.38	0.97	0.70	0.33	0.97	0.71
	RF	Training	0.26	0.98	0.79	0.27	0.98	0.86
		Test	0.60	0.90	1.27	0.54	0.91	1.24
	XGBoost	Training	0.15	1.00	0.20	0.19	1.00	0.25
		Test	0.45	0.94	0.97	0.42	0.96	0.82
	RF-GBM	Training	0.16	1.00	0.21	0.08	1.00	0.10
		Test	0.40	0.97	0.71	0.41	0.96	0.83

3.3.1.1 Train-Test Split (90:10)

ANN model achieved nearly perfect results for BOD_e on the training set (MAE of 0.09, R^2 of 1.00, and RMSE of 0.14) and maintained strong performance on the test set (MAE of 0.60, R^2 of 0.96, and RMSE of 1.25). The GBM model had slight errors on the test set compared to ANN. The RF model showed higher errors on the test set compared to other models. XGBoost and RF-GBM maintained good performance on the test set. For $(NH_3)_e$, the ANN model had good training performance but poor test set performance (MAE of 0.53, R^2 of 0.33, and RMSE of 1.06). The GBM model showed better set results (MAE of 0.48, R^2 of 0.60, and RMSE of 0.82). RF, XGBoost and RF-GBM showed similar results as GBM. For P_e, the ANN model had reasonable training performance, but poor test set performance (test MAE of 0.14, R^2 of 0.42, and RMSE of 0.18). The GBM model had better test set results (MAE of 0.10, R^2 of 0.65, and RMSE of 0.14). The RF model had moderate performance with slightly higher test set errors (MAE of 0.11, R^2 of 0.61, and RMSE of 0.15). XGBoost and RF-GBM showed good performance on both sets. For TSS_e, the ANN model showed excellent training performance and strong test

set results (test MAE of 0.52, R^2 of 0.95, and RMSE of 0.91). The GBM model had better test set results (MAE of 0.38, R^2 of 0.97, and RMSE of 0.70). The RF model showed moderate performance with higher test set errors (MAE of 0.60, R^2 of 0.90, and RMSE of 1.27). XGBoost and RF-GBM maintained good performance for TSS_e.

3.3.1.2 Train-Test Split (80:20)

For BOD_e, ANN model achieved almost perfect results on the training set (MAE of 0.09, R^2 of 1.00, and RMSE of 0.14) and maintained strong performance on the test set (MAE of 0.60, R^2 of 0.96, and RMSE of 1.25). The GBM model exhibited slight errors on the test set (MAE of 0.68, R^2 of 0.94, and RMSE of 1.35). The RF model had higher errors on the test set compared to other models. The XGBoost and RF-GBM models maintained good performance on the test set.

For $(NH_3)_e$, ANN model had good training performance (MAE of 0.11, R^2 of 0.99, and RMSE of 0.16) but poor test set performance (MAE of 0.46, R^2 of 0.35, and RMSE of 0.89). The GBM model had better test set results (MAE of 0.47, R^2 of 0.45, and RMSE of 0.82). The RF, XGBoost, and RF-GBM models had good performance on the test set.

For P_e, ANN model had reasonable training and test performance. The GBM model had better test set results (test MAE of 0.11, R^2 of 0.55, and RMSE of 0.17). The RF, XGBoost and RF-GBM models had good performance on both sets.

For TSS_e, ANN model had excellent training performance and strong test set results (MAE of 0.56, R^2 of 0.94, and RMSE of 0.96). The GBM model had better test set results (MAE of 0.33, R^2 of 0.97, and RMSE of 0.71). The RF model had moderate performance with higher test set errors (MAE of 0.54, R^2 of 0.91, and RMSE of 1.24). The XGBoost and RF-GBM models maintained good performance on both sets (MAE of 0.42, R^2 of 0.96, and RMSE of 0.82).

3.3.2 Feature Selection Methods

Various FS methods were employed to identify the most significant variables impacting the concentrations of BOD_e, $(NH_3)_e$, P_e, and TSS_e in WWTP. Table 3.3 shows common features shared by FS methods for various target variables. For BOD_e, TSS_e is identified as the most significant variable across multiple methods, with BOD Removed (%) frequently highlighted as important. For $(NH_3)_e$, $Flow_i$ is most significant across methods and BOD_e is constantly significant in all methods. For P_e, TSS_e and BOD_e are topmost and second most across all the methods. For TSS_e, BOD and TSS Removed (%) are most significant in all methods. The consistency across different FS methods strengthens the reliability of these findings providing a robust basis for further research and practical applications.

Table 3.3. Common features selected by FS methods.

No. of features	Name of features	Target variable
3	BOD Removed (%), P_e, $(NH_3)_e$	BOD_e
2	BOD_e, P_e	$(NH_3)_e$
2	BOD_e, TSS_e	P_e
2	BOD Removed (%), TSS Removed (%)	TSS_e

3.3.3 XAI

The results of the LIME and SHAP analyses for various target variables revealed the order of feature influence and their effects on ML models. Figure 3.2 shows one of the LIME plots. The figure shows the variables and their contributions (blue as negative, orange as positive) to BOD_e for RF-GBM model for 50th instance. Predicted values of 50th instance for BOD_e is 15.29 mg/L. According to the figure, TSS_e show strongest positive effect on BOD_e prediction. Figure 3.3 shows one of the SHAP summary plots that demonstrates variables and

their contributions to BOD_e for RF-GBM model. The figure shows that higher values of TSS_i (red dots) tend to contribute positively to the BOD_i prediction. In comparison, the lower values (blue dots) have negative contributions.

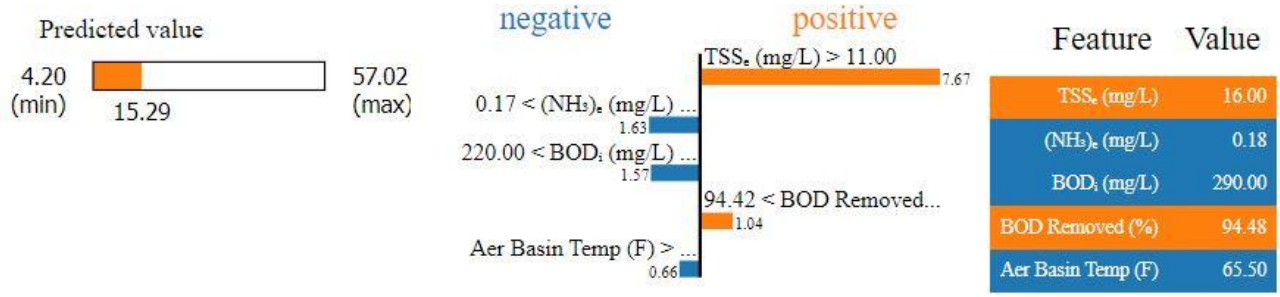

Figure 3.2. LIME explanation for RF-GBM model for BOD_e; LIME predicted 15.29 with a range between 4.20 and 57.02. TSS_e with a value of 16.00 mg/L is the most significant feature positively influencing the BOD_e prediction. $(NH_3)_e$, BOD_i, and Aer Basin Temp negatively affect the prediction.

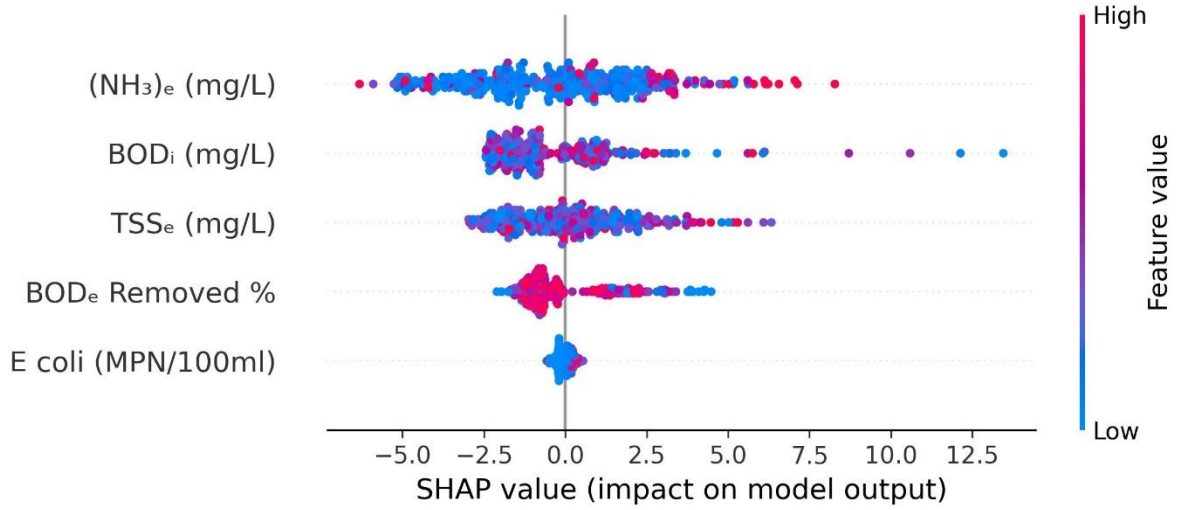

Figure 3.3. SHAP explanation for RF-GBM model (BOD_e).

Multiple variables were shared by both LIME and SHAP. Since LIME provided positive and negative impacts of variables explicitly, signs (positive or negative) were provided next to

the variable's name. The SHAP summary plot was indecisive regarding positive or negative influence in many cases. Therefore, no signs were provided next to the SHAP identified variables. For BOD_e, LIME and SHAP analyses consistently identified TSS_e and BOD Removed (%) as influential variables across all models, with some discrepancies in the order of influence. For TSS_e, both LIME and SHAP analyses indicated TSS Removed (%), BOD_e, and P_e as significant variables. For $(NH_3)_e$, LIME consistently highlighted BOD_e and E. Coli as significant variables, with varying impact directions across the models. SHAP's results were less consistent, with BOD_e and E. Coli appearing in differing orders of importance. For P_e, LIME and SHAP both identified TSS_e, BOD_e, $Temp_e$ and Aer Basin Temp as important features, with varying order or influence.

3.4 Discussion

The study investigated the performance of multiple ML models, i.e., ANN, GBM, RF, XGBoost, and RF-GBM, in predicting several influential influent and effluent water quality variables in a WWTP. ANN, GBM, and XGBoost demonstrated significant potential for variable prediction as they produced low error rates and strong correlation coefficients (R^2).

Based on the findings of the study, the complex interactions among various WWTP variables can be captured by GBM. For example, GBM performed particularly well in predicting variables such as BOD and NH_3. This agrees with other study that GBM performed better than ANN in WWTP variable prediction (Bagherzadeh et al., 2021). Although RF performed very well on training data, overfitting caused poor performance on the test set (unseen data). As an alternative to RF model, hybrid RF-GBM model was able to increase the models' accuracy particularly for predicting BOD and P levels, by utilizing the advantages of both models. Overall,

hybrid RF-GBM model provided a flexible approach that can be tailored to specific prediction challenges within WWTPs. ANN provided a competitive alternative, while GBM, XGBoost, and RF-GBM stood out as superior performers. The performance of XGBoost is consistent with other researchers' findings (Shao et al., 2023; Zhang et al., 2024). XGBoost utilizes gradient-boosting methods to sequentially create an ensemble of weak prediction models and fix errors, leading to greater overall performance (Chen and Guestrin 2016).

The LIME and SHAP analyses produced strong agreement with the FS results. Table 3.4 compares the shared feature(s) chosen by the FS methods with the features chosen by LIME and SHAP for the ML models. Traditionally, FS methods are used in various studies to identify the most suitable input data from a dataset to increase model accuracy (Bagherzadeh et al., 2021). While FS methods do not consider ML models in selecting influential variables for target variables, XAI tools i.e. LIME and SHAP show influential variables significance on each models' prediction. The study revealed that FS and XAI have identified several common influential variables regardless of choice of model or FS methods in predicting target variables.

Table 3.4. Comparison of the shared feature(s) chosen by the FS methods with the features chosen by LIME and SHAP.

Target variable	Common features by FS methods	LIME	SHAP
BOD_e	BOD Removed (%)	ANN, GBM, RF, RF-GBM	GBM, RF, RF-GBM, XGBoost
	$(NH_3)_e$	GBM, RF, RF-GBM, XGBoost	ANN, GBM, RF, RF-GBM, XGBoost
	P_e	RF	RF
$(NH_3)_e$	BOD_e	ANN, GBM, RF, RF-GBM, XGBoost	GBM, RF, RF-GBM, XGBoost
	P_e	ANN, GBM, RF-GBM, XGBoost	-
P_e	BOD_e	ANN, GBM, RF, RF-GBM, XGBoost	GBM, RF, RF-GBM, XGBoost
	TSS_e	ANN, GBM, RF, RF-GBM, XGBoost	GBM, RF, RF-GBM, XGBoost
TSS_e	BOD Removed (%)	-	XGBoost
	TSS Removed (%)	ANN, GBM, RF, RF-GBM, XGBoost	ANN, GBM, RF, RF-GBM, XGBoost

It is also interesting to find that although ML models perform without knowledge of real-world impact of input variables on target variables, some of the common variables significantly impact certain models according to XAI. For instance, BOD_e, TSS_i, and P_i were all shown to be significant to BOD_i predictions by both LIME and SHAP. LIME explicitly reported positive and negative impacts while SHAP summary plot displayed varying importance without an apparent direction of influence. Based on the findings, LIME and SHAP can help in understanding the variables' importance in ML-based prediction, thereby can support targeted interventions in WWTP operation.

3.5 Conclusions

This study compared several XAI tools in predicting key WWTP variables using various ML models. Based on the findings of this study, ML models, ANN, GBM, XGBoost, and RF-GBM consistently outperform the others, exhibiting strong prediction abilities with reduced errors and higher R^2 values. The use of SHAP and LIME enhances the interpretability of ML models by providing the impact of input variables on the model outputs. The reliability of XAI tools in identifying important WWTP factors is supported by the agreement of results between FS approaches and XAI tools. The effects of various variable sets on model performance or dimension reduction strategies can also be further investigated. WWTP can optimize operations and reduce costs while mitigating environmental impacts by leveraging the interpretation provided by XAI and using robust ML models.

3.6 References

1. Torregrossa, D., Schutz, G., Cornelissen, A., Hernández-Sancho, F., & Hansen, J. (2016). Energy saving in WWTP: Daily benchmarking under uncertainty and data availability limitations. Environmental Research, 148, 330–337.

2. Abba, S. I., & Elkiran, G. (2017). Effluent prediction of chemical oxygen demand from the wastewater treatment plant using artificial neural network application. Procedia Computer Science, 120, 156–163.

3. Bernardelli, A., Marsili-Libelli, S., Manzini, A., Stancari, S., Tardini, G., Montanari, D., et al. (2020). Real-time model predictive control of a wastewater treatment plant based on machine learning. Water Science and Technology, 81, 2391–2400.

4. Zhang, S., Wang, H., & Keller, A. A. (2021). Novel machine learning-based energy consumption model of wastewater treatment plants. ACS ES&T Water, 1, 2531–2540.

5. Guo, H., Jeong, K., Lim, J., Jo, J., Kim, Y. M., Park, J. P., ... & Cho, K. H. (2015). Prediction of effluent concentration in a wastewater treatment plant using machine learning models. Journal of Environmental Sciences, 32, 90-101.

6. Wang, D., Thunéll, S., Lindberg, U., Jiang, L., Trygg, J., Tysklind, M., & Souihi, N. (2021). A machine learning framework to improve effluent quality control in wastewater treatment plants. Science of the total environment, 784, 147138.

7. El-Rawy, M., Abd-Ellah, M. K., Fathi, H., & Ahmed, A. K. A. (2021). Forecasting effluent and performance of wastewater treatment plant using different machine learning techniques. Journal of Water Process Engineering, 44, 102380.

8. Li, G., Ji, J., Ni, J., Wang, S., Guo, Y., Hu, Y., ... & Li, Y. Y. (2022). Application of deep learning for predicting the treatment performance of real municipal wastewater based on one-year operation of two anaerobic membrane bioreactors. Science of the Total Environment, 813, 151920.

10. Zhu, J. J., Borzooei, S., Sun, J., & Ren, Z. J. (2022). Deep learning optimization for soft sensing of hard-to-measure wastewater key variables. ACS ES&T Engineering, 2(7), 1341-1355.

11. Aghdam, E., Mohandes, S. R., Manu, P., Cheung, C., Yunusa-Kaltungo, A., & Zayed, T. (2023). Predicting quality parameters of wastewater treatment plants using artificial intelligence techniques. Journal of Cleaner Production, 405, 137019.

12. Shyu, H. Y., Castro, C. J., Bair, R. A., Lu, Q., & Yeh, D. H. (2023). Development of a soft sensor using machine learning algorithms for predicting the water quality of an onsite wastewater treatment system. ACS Environmental Au, 3(5), 308-318.

13. Wei, X., Yu, J., Tian, Y., Ben, Y., Cai, Z., & Zheng, C. (2023). Comparative performance of three machine learning models in predicting influent flow rates and nutrient loads at wastewater treatment plants. ACS ES&T Water, 4(3), 1024-1035.

14. Xu, Y., Wang, Z., Nairat, S., Zhou, J., & He, Z. (2023). Artificial intelligence-assisted prediction of effluent phosphorus in a full-scale wastewater treatment plant with missing phosphorus input and removal data. ACS ES&T Water, 4(3), 880-889.

15. Yu, J., Tian, Y., Jing, H., Sun, T., Wang, X., Andrews, C. B., & Zheng, C. (2023). Predicting regional wastewater treatment plant discharges using machine learning and population migration big data. ACS ES&T Water, 3(5), 1314-1328.

16. Alsulaili, A., & Refaie, A. (2021). Artificial neural network modeling approach for the prediction of five-day biological oxygen demand and wastewater treatment plant performance. Water Supply, 21(5), 1861-1877.

17. Fan, M., Hu, J., Cao, R., Ruan, W., & Wei, X. (2018). A review on experimental design for pollutants removal in water treatment with the aid of artificial intelligence. Chemosphere, 200, 330-343.

18. Adibimanesh, B., Polesek-Karczewska, S., Bagherzadeh, F., Szczuko, P., & Shafighfard, T. (2023). Energy consumption optimization in wastewater treatment plants: Machine learning for monitoring incineration of sewage sludge. Sustainable energy technologies and assessments, 56, 103040.

19. Keerio, H. A., Shah, S. A., Ali, Z., Panhwar, S., Solangi, G. S., Ali, A., ... & Yong, Y. C. (2024). A fascinating exploration into nitrite accumulation into low concentration reactors using cutting-edge machine learning techniques. Process Biochemistry, 146, 160-168.

20. Solangi, G. S., Ali, Z., Bilal, M., Junaid, M., Panhwar, S., Keerio, H. A., ... & Zaman, N. (2024). Machine learning, Water Quality Index, and GIS-based analysis of groundwater quality. Water Practice & Technology, 19(2), 384-400.

21. Bagherzadeh, F., Mehrani, M. J., Basirifard, M., & Roostaei, J. (2021). Comparative study on total nitrogen prediction in wastewater treatment plant and effect of various feature selection methods on machine learning algorithms performance. Journal of Water Process Engineering, 41, 102033.

22. Wu, Z., Duan, H., Li, K., & Ye, L. (2022). A comprehensive carbon footprint analysis of different wastewater treatment plant configurations. Environmental Research, 214, 113818.

23. Keerio, H. A., Bae, W., Park, J., & Kim, M. (2020). Substrate uptake, loss, and reserve in ammonia-oxidizing bacteria (AOB) under different substrate availabilities. Process Biochemistry, 91, 303-310.

24. Singh, N. K., Yadav, M., Singh, V., Padhiyar, H., Kumar, V., Bhatia, S. K., & Show, P. L. (2023). Artificial intelligence and machine learning-based monitoring and design of biological wastewater treatment systems. Bioresource technology, 369, 128486.

25. Zhao, L. J., Chai, T. Y., & Yuan, D. C. (2012). Selective ensemble extreme learning machine modeling of effluent quality in wastewater treatment plants. International Journal of Automation and Computing, 9(6), 627-633.

26. Bagheri, M., Mirbagheri, S. A., Ehteshami, M., Bagheri, Z., & Kamarkhani, A. M. (2016). Analysis of variables affecting mixed liquor volatile suspended solids and prediction of effluent quality parameters in a real wastewater treatment plant. Desalination and Water Treatment, 57(45), 21377-21390.

27. Sharghi, E., Nourani, V., AliAshrafi, A., & Gökçekuş, H. (2019). Monitoring effluent quality of wastewater treatment plant by clustering based artificial neural network method. Desalination and Water Treatment, 164, 86-97.

28. Khatri, N., Khatri, K. K., & Sharma, A. (2019). Prediction of effluent quality in ICEAS-sequential batch reactor using feedforward artificial neural network. Water science and technology, 80(2), 213-222.

29. Al-Ghazawi, Z., & Alawneh, R. (2021). Use of artificial neural network for predicting effluent quality parameters and enabling wastewater reuse for climate change resilience–A case from Jordan. Journal of Water Process Engineering, 44, 102423.

30. Elmaadawy, K., Abd Elaziz, M., Elsheikh, A. H., Moawad, A., Liu, B., & Lu, S. (2021). Utilization of random vector functional link integrated with manta ray foraging optimization for effluent prediction of wastewater treatment plant. Journal of Environmental Management, 298, 113520.

31. Nourani, V., Asghari, P., & Sharghi, E. (2021). Artificial intelligence based ensemble modeling of wastewater treatment plant using jittered data. Journal of Cleaner Production, 291, 125772.

32. Ly, Q. V., Truong, V. H., Ji, B., Nguyen, X. C., Cho, K. H., Ngo, H. H., & Zhang, Z. (2022). Exploring potential machine learning application based on big data for prediction of wastewater quality from different full-scale wastewater treatment plants. Science of the Total Environment, 832, 154930.

33. Dantas, M. S., Christofaro, C., & Oliveira, S. C. (2023). Artificial neural networks for performance prediction of full-scale wastewater treatment plants: a systematic review. Water Science & Technology, 88(6), 1447-1470.

34. Mahanna, H., El-Rashidy, N., Kaloop, M. R., El-Sapakh, S., Alluqmani, A., & Hassan, R. (2024). Prediction of wastewater treatment plant performance through machine learning techniques. Desalination and Water Treatment, 319, 100524.

35. Park, J., Lee, W. H., Kim, K. T., Park, C. Y., Lee, S., & Heo, T. Y. (2022). Interpretation of ensemble learning to predict water quality using explainable artificial intelligence. Science of the Total Environment, 832, 155070.

36. Hu, Y., Wei, R., Yu, K., Liu, Z., Zhou, Q., Zhang, M., ... & Qu, S. (2024). Exploring sludge yield patterns through interpretable machine learning models in China's municipal wastewater treatment plants. Resources, Conservation and Recycling, 204, 107467.

37. Shao, S., Fu, D., Yang, T., Mu, H., Gao, Q., & Zhang, Y. (2023). Analysis of machine learning models for wastewater treatment plant sludge output prediction. Sustainability, 15(18), 13380.

38. Shafighfard, T., Kazemi, F., Asgarkhani, N., & Yoo, D. Y. (2024). Machine-learning methods for estimating compressive strength of high-performance alkali-activated concrete. Engineering Applications of Artificial Intelligence, 136, 109053.

30. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

40. Lundberg, S. M., Erion, G. G., & Lee, S. I. (2018). Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888.

41. Li, R., Feng, K., An, T., Cheng, P., Wei, L., Zhao, Z., ... & Zhu, L. (2024). Enhanced insights into effluent prediction in wastewater treatment plants: Comprehensive deep learning model explanation based on shap. ACS ES&T Water, 4(4), 1904-1915.

42. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).

43. Ye, Z., Yang, J., Zhong, N., Tu, X., Jia, J., & Wang, J. (2020). Tackling environmental challenges in pollution controls using artificial intelligence: A review. Science of the Total Environment, 699, 134279.

44. Matheri, A. N., Ntuli, F., Ngila, J. C., Seodigeng, T., & Zvinowanda, C. (2021). Performance prediction of trace metals and cod in wastewater treatment using artificial neural network. Computers & Chemical Engineering, 149, 107308.

45. Konstantinov, A. V., & Utkin, L. V. (2021). Interpretable machine learning with an ensemble of gradient boosting machines. Knowledge-Based Systems, 222, 106993.

46. Tyralis, H., Papacharalampous, G., & Langousis, A. (2019). A brief review of random forests for water scientists and practitioners and their recent history in water resources. Water, 11(5), 910.

47. Nafsin, N., & Li, J. (2023). Prediction of total organic carbon and E. coli in rivers within the Milwaukee River basin using machine learning methods. Environmental Science: Advances, 2(2), 278-293.

48. Jiang, M., Wang, J., Hu, L., & He, Z. (2023). Random forest clustering for discrete sequences. Pattern Recognition Letters, 174, 145-151.

49. Szomolányi, O., & Clement, A. (2023). Use of random forest for assessing the effect of water quality parameters on the biological status of surface waters. GEM-International Journal on Geomathematics, 14(1), 20.

50. Sun, Z., Wang, G., Li, P., Wang, H., Zhang, M., & Liang, X. (2024). An improved random forest based on the classification accuracy and correlation measurement of decision trees. Expert Systems with Applications, 237, 121549.

51. Yadav, P., Chandra, M., Fatima, N., Sarwar, S., Chaudhary, A., Saurabh, K., & Yadav, B. S. (2023). Predicting influent and effluent quality parameters for a UASB-based wastewater treatment plant in Asia covering data variations during COVID-19: A machine learning approach. Water, 15(4), 710.

52. Xie, Y., Chen, Y., Lian, Q., Yin, H., Peng, J., Sheng, M., & Wang, Y. (2022). Enhancing real-time prediction of effluent water quality of wastewater treatment plant based on improved feedforward neural network coupled with optimization algorithm. Water, 14(7), 1053.

53. Sargent, R. G. (2010, December). Verification and validation of simulation models. In Proceedings of the 2010 winter simulation conference (pp. 166-183). IEEE.

54. Tsioptsias, N., Tako, A., & Robinson, S. (2016). Model validation and testing in simulation: a literature review. In 5th student conference on operational research (SCOR 2016) (pp. 6-1). Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

55. Zhang, X., & Liu, C. A. (2023). Model averaging prediction by K-fold cross-validation. Journal of Econometrics, 235(1), 280-301.

56. Kazemi, F., Asgarkhani, N., Shafighfard, T., Jankowski, R., & Yoo, D. Y. (2025). Machine-learning methods for estimating performance of structural concrete members reinforced with fiber-reinforced polymers. Archives of Computational Methods in Engineering, 32(1), 571-603.

57. Bagherzadeh, F., Shafighfard, T., Khan, R. M. A., Szczuko, P., & Mieloszyk, M. (2023). Prediction of maximum tensile stress in plain-weave composite laminates with interacting holes via stacked machine learning algorithms: A comparative study. Mechanical Systems and Signal Processing, 195, 110315.

58. Shafighfard, T., Bagherzadeh, F., Rizi, R. A., & Yoo, D. Y. (2022). Data-driven compressive strength prediction of steel fiber reinforced concrete (SFRC) subjected to elevated temperatures using stacked machine learning algorithms. Journal of materials research and technology, 21, 3777-3794.

59. Safder, U., Kim, J., Pak, G., Rhee, G., & You, K. (2022). Investigating machine learning applications for effective real-time water quality parameter monitoring in full-scale wastewater treatment plants. Water, 14(19), 3147.

60. Zhang, Y., Wu, H., Xu, R., Wang, Y., Chen, L., & Wei, C. (2024). Machine learning modeling for the prediction of phosphorus and nitrogen removal efficiency and screening of crucial microorganisms in wastewater treatment plants. Science of The Total Environment, 907, 167730.

61. Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

CHAPTER 4: APPLICATION OF EXPLAINABLE ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING IN PREDICTING WASTEWATER TREATMENT PLANT VARIABLES: A COMPARATIVE STUDY OF SMALL AND LARGE-SCALE TREATMENT PLANTS

4.1 Introductions

Wastewater treatment plants (WWTPs) play a crucial role in sustainable water management by removing pollutants from sewage before it is discharged into the environment. Efficient WWTP operations ensure regulatory standards, cost-effective treatment, and sustainable urban development. However, wastewater treatment processes are governed by nonlinear, multivariate, and dynamic interactions among physical, chemical, and biological variables, posing significant challenges for accurate monitoring, control, and prediction (Arismendy et al., 2020; Alvi et al., 2023; Newhart et al., 2023).

Traditionally, WWTPs have depended on empirical models and manual monitoring to support decision making and process optimization. However, conventional approaches often fail to capture the intricate relationships between the influent characteristics, process variables, and effluent quality. Recently, machine learning (ML) has emerged as a powerful tool for analyzing complex environmental systems, including WWTPs. ML techniques demonstrate significant accuracy in predicting key variables, such as ammonia nitrogen (NH_3-N), biochemical oxygen demand (BOD), chemical oxygen demand (COD), total phosphorus (TP), and total suspended solids (TSS) (El-Rawy et l., 2021; Wang et al., 2021; Zhang et al., 2021; Zhu et al., 2022; Aghdam et al., 2023; Shyu et al., 2023; Wei et al., 2023; Yu et al., 2023; Cechinel et al., 2024; Xu et al., 2024). These data-driven models are capable of capturing hidden patterns in operational data, enabling WWTPs to optimize treatment processes, improve effluent quality, and reduce operational costs.

Although ML has proven to be successful in predicting wastewater variables with high accuracy, WWTP operators cannot rely solely on black-box ML models that lack transparency. This lack of interpretability is a major barrier to its practical adoption in WWTP operations, where operators and regulators require a clear understanding of the factors that influence model outputs for process control, regulatory compliance, and risk management. Explainable artificial intelligence (XAI) has recently gained attention as a means of overcoming this challenge by enhancing the transparency and interpretability of ML models, and providing insights into how and why specific outcomes are generated (Park et al., 2022; Hu et al., 2024). However, most studies have focused on individual facilities that have relatively stable operating conditions. A significant knowledge gap exists regarding the performance of ML models and interpretability techniques across WWTPs of various sizes. Small-scale WWTPs are particularly susceptible to fluctuations in influent loads, limited monitoring infrastructure, and inconsistent operational practices, while large-scale WWTPs tend to exhibit consistent data collection and process stability. A comparative analysis of ML performance across both small and large facilities can provide valuable insights into the scalability and adaptability of these modeling approaches. Understanding the scalability and robustness of ML and XAI tools across different WWTP sizes is essential for promoting their broader application in the wastewater sector.

This study investigated the application of ML and XAI in predicting critical effluent quality variables at WWTPs and compared their effectiveness in both small- and large-scale facilities. This study employed Xtreme Gradient Boosting (XGBoost) and Random Forests (RF) to investigate the performance of ML algorithms in WWTPs with varying data availability. Additionally, XAI techniques, SHapley Additive explanations (SHAP), and Local Interpretable

Model-agnostic Explanations (LIME) were employed to interpret predictions, providing a deeper understanding of the key factors influencing the performance of the ML model. The findings of this study contribute to bridging the gap between advanced data-driven modeling and real-world implementation in diverse WWTP settings, thereby offering guidance for integrating interpretable ML models into wastewater treatment operations.

4.2 Methods

4.2.1 Data collection and processing

Data representing a range of treatment capacities were collected from four plants in Wisconsin, USA. The facilities include: Monroe Wastewater Treatment Facility (WWTF) with a capacity of 3.7 million gallons per day (MGD), Sheboygan WWTF with an average capacity of 18.4 MGD and a peak design capacity of 58.6 MGD, Madison Metropolitan Sewerage District (MMSD) with an approximate capacity of 37 MGD of wastewater to the plant, and Milwaukee Metropolitan Sewerage District South Shore Water Reclamation Facility (MMSD SSWRF) with a capacity to treat 250 MGD daily flow and 300 MGD peak hourly flow. The selection of two smaller and two larger treatment plants allows for the comparison of ML and XAI performances in facilities of different scales. The dataset consists of daily observations from January 1, 2019, to November 30, 2024. Figures4.1, 4.2, 4.3, and 4.4 and Tables 4.1, 4.2, 4.3, and 4.4 showed variable distribution and statistics for Monroe, Sheboygan, MMSD SSWRF, and MMSD, respectively.

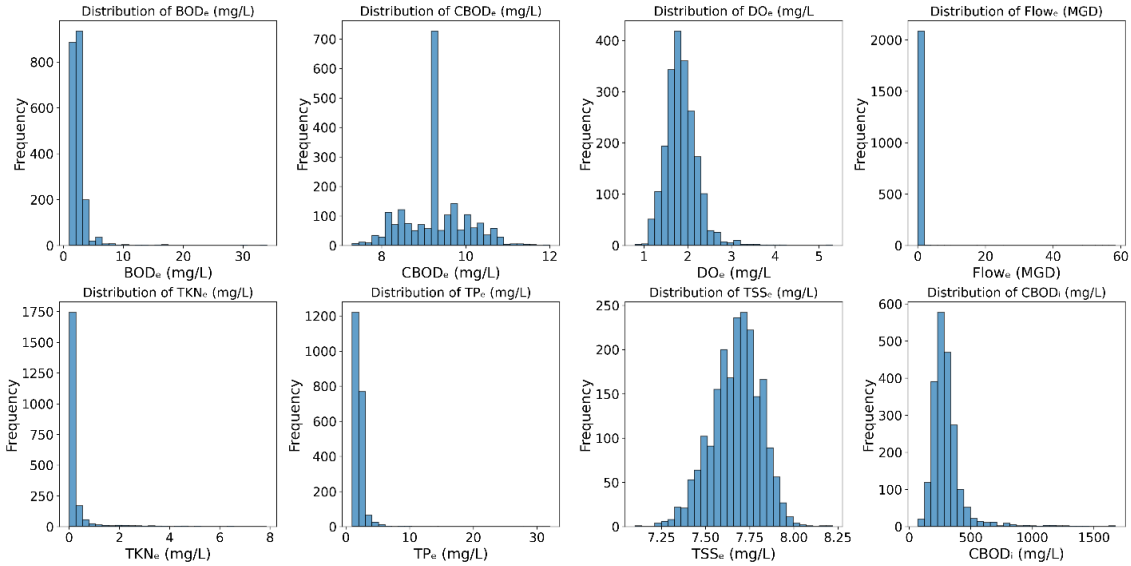

Figure 4.1. Monroe WWTF variable distribution

Table 4.1. Monroe WWTF variable statistics

Variable Name	Minimum	Maximum	Mean	Standard Deviation	Coefficient of Variation
BOD_e (mg/L)	1.00	34.00	2.80	1.50	0.54
$CBOD_e$ (mg/L)	7.30	12.00	9.30	0.73	0.08
DO_e (mg/L	0.80	5.31	1.86	0.36	0.19
$Flow_e$ (MGD)	0.04	58.40	0.33	2.60	7.87
TKN_e (mg/L)	0.01	7.88	0.30	0.71	2.39
TP_e (mg/L)	1.00	32.00	2.06	1.25	0.61
TSS_e (mg/L)	7.10	8.22	7.67	0.14	0.02
$CBOD_i$ (mg/L)	70.00	1675.00	302.67	137.00	0.45

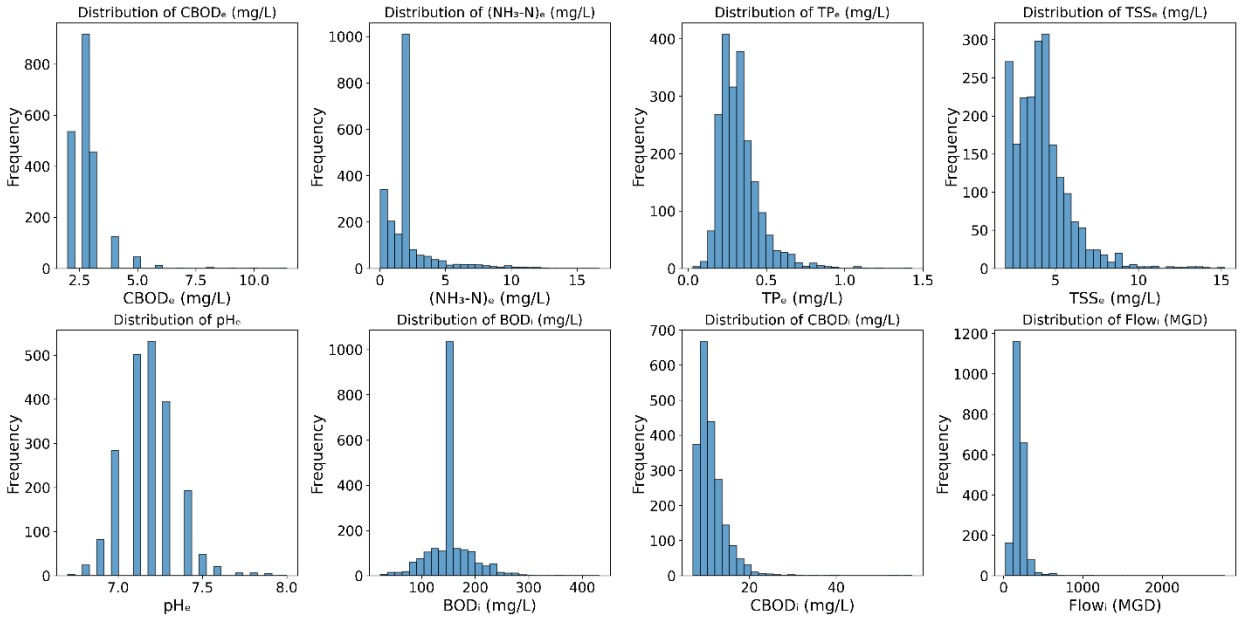

Figure 4.2. Sheboygan WWTF variable distribution

Table 4.2. Sheboygan WWTF variable statistics

Variable Name	Minimum	Maximum	Mean	Standard Deviation	Coefficient of Variation
$CBOD_e$ (mg/L)	2.00	11.40	2.81	0.77	0.28
$(NH_3\text{-}N)_e$ (mg/L)	0.04	16.70	2.23	1.89	0.85
TP_e (mg/L)	0.03	1.43	0.32	0.13	0.40
TSS_e (mg/L)	2.00	15.20	4.16	1.62	0.39
pH_e	6.70	8.00	7.19	0.16	0.02
BOD_i (mg/L)	23.00	432.00	155.09	38.45	0.25
$CBOD_i$ (mg/L)	7.06	57.44	11.33	3.61	0.32
$Flow_i$ (MGD)	25.00	2790.00	204.05	107.17	0.53

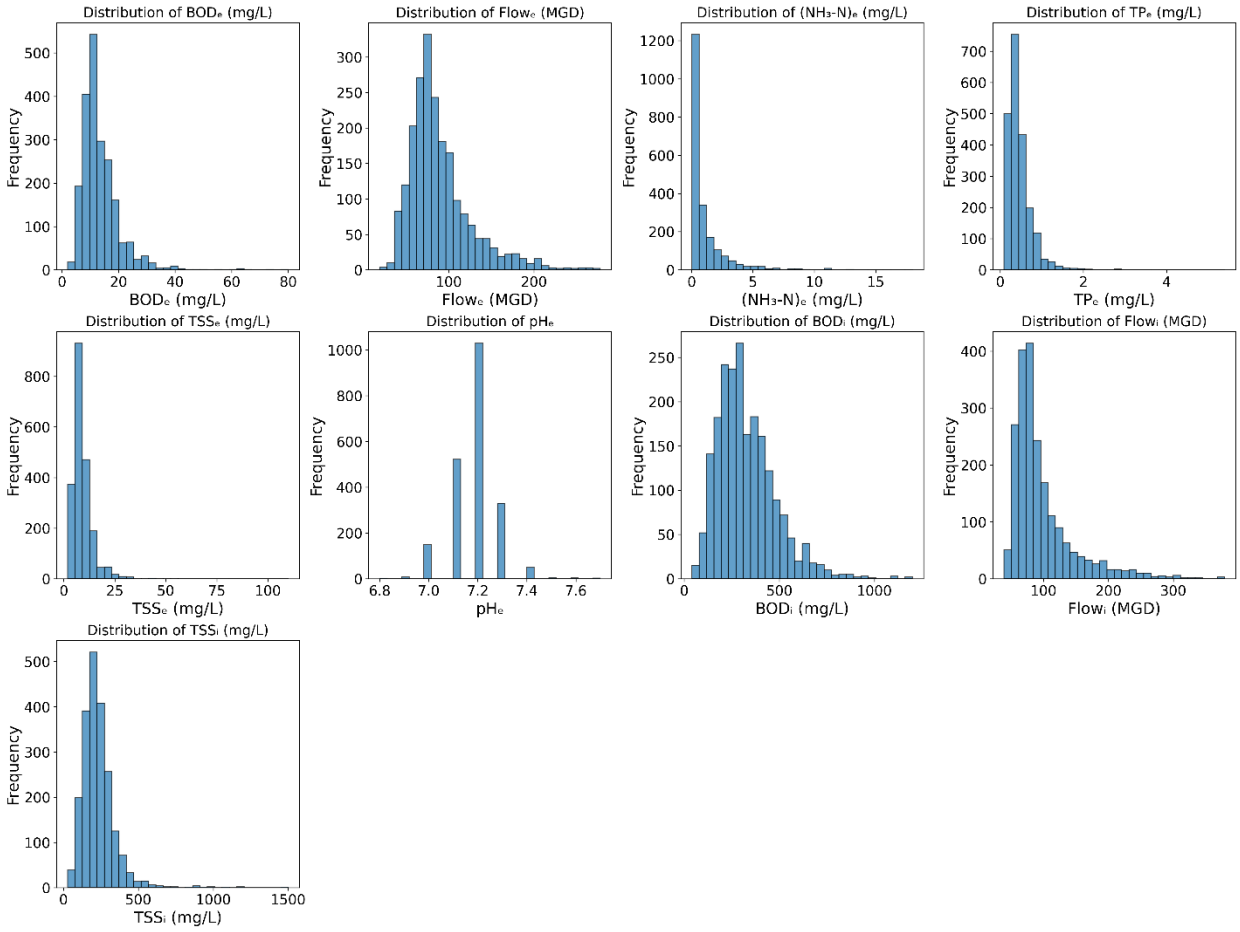

Figure 4.3. MMSD SSWRF variable distribution

Table 4.3. MMSD SSWRF variable statistics

Variable Name	Minimum	Maximum	Mean	Standard Deviation	Coefficient of Variation
BOD_e (mg/L)	2.00	80.00	13.42	6.86	0.51
$Flow_e$ (MGD)	19.00	278.00	89.52	36.90	0.41
$(NH_3\text{-}N)_e$ (mg/L)	0.02	18.00	1.08	1.67	1.55
TP_e (mg/L)	0.09	5.40	0.46	0.31	0.68
TSS_e (mg/L)	1.90	110.00	9.32	6.06	0.65
pH_e	6.80	7.70	7.18	0.09	0.01
BOD_i (mg/L)	40.00	1200.00	326.89	152.79	0.47
$Flow_i$ (MGD)	39.00	379.00	97.41	48.04	0.49
TSS_i (mg/L)	28.00	1500.00	232.17	113.13	0.49

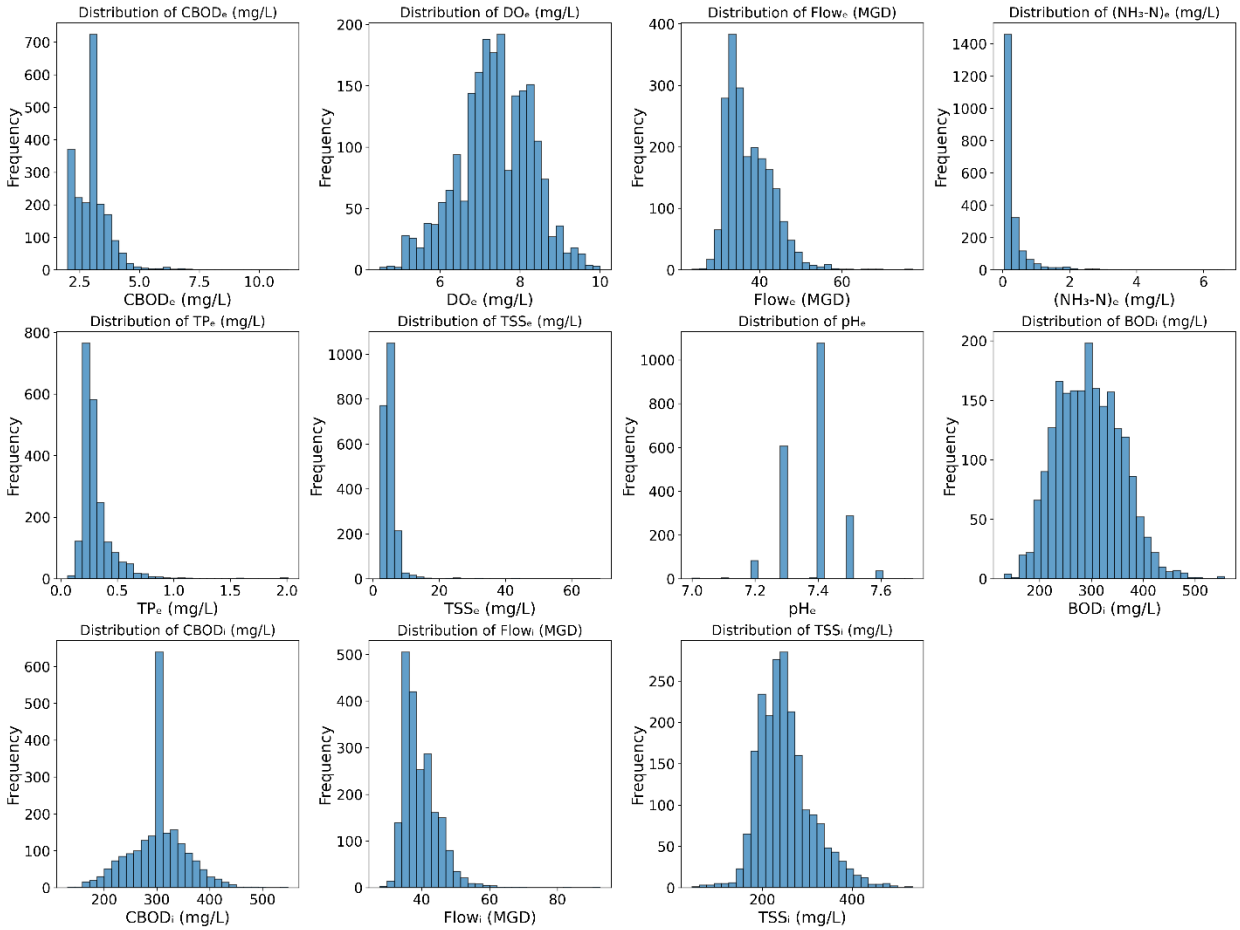

Figure 4.4. MMSD variable distribution

Table 4.4. MMSD variable statistics

Variable Name	Minimum	Maximum	Mean	Standard Deviation	Coefficient of Variation
$CBOD_e$ (mg/L)	2.00	11.20	3.04	0.77	0.25
DO_e (mg/L)	4.50	10.00	7.37	0.92	0.12
$Flow_e$ (MGD)	23.57	77.12	37.51	5.53	0.15
$(NH_3\text{-}N)_e$ (mg/L)	0.05	6.63	0.32	0.52	1.62
TP_e (mg/L)	0.06	2.01	0.31	0.16	0.52
TSS_e (mg/L)	2.10	68.60	5.19	2.83	0.55
pH_e	7.00	7.70	7.38	0.08	0.01
BOD_i (mg/L)	132.00	557.00	295.64	61.60	0.21
$CBOD_i$ (mg/L)	131.00	548.00	301.76	51.73	0.17
$Flow_i$ (MGD)	27.71	92.55	39.67	5.25	0.13
TSS_i (mg/L)	44.20	534.00	251.51	60.53	0.24

Data were thoroughly examined for missing entries. If more than 50% of the values for a given variable were missing, that variable was removed from the analysis. Thus, missing values were replaced with the mean of the available values for variables with < 50% of missing data. To reduce multicollinearity, no two variables with an absolute Pearson correlation coefficient ≥ 0.9 were used together (one of the pairs was removed). In this study, the effluent quality parameters (NH_3-N, BOD, COD, TP, and TSS) were treated as the target variables to be predicted, and various influent and operational parameters served as input features. For clarity, we use the subscript "$_i$" to denote influent variables and "$_e$" to denote effluent variables (e.g., BOD_i is influent BOD, and BOD_e is effluent BOD).

4.2.2 Feature selection (FS)

A suite of well-known feature selection (FS) methods was applied to identify influential input variables for each target. Specifically, we evaluated the least absolute shrinkage and selection operator (LASSO), mutual information (MI), random forest feature importance, Pearson correlation, and principal component analysis (PCA) for their ability to rank the predictor importance. As these methods produce important scores on different scales, normalized importance scaling is used to standardize all scores to the range [0, 1], where 0 corresponds to the least important feature and 1 corresponds to the most important. This normalization enabled a fair comparison of feature importance across different selection techniques.

4.2.3 SHAP

SHAP is a game-theoretic XAI method that explains model predictions by attributing each prediction to the contributions of individual features (Lundberg and Lee, 2017; Lundberg et al., 2018; Huang et al., 2020; Shao et al., 2023; Li et al., 2024; Xu et al., 2024). SHAP values are computed based on Shapley values from cooperative game theory, effectively assessing all possible feature-value combinations to determine how the presence or absence of each feature impacts the model outcome. SHAP provides both global and local explanations. Global SHAP values summarize the feature importance across the entire dataset, whereas local SHAP values explain the contribution of features to individual predictions. In this study, SHAP analysis was used to understand the relative impact of each input variable on the model output and to identify the features that were most influential in driving the overall predictions.

4.2.4 LIME

LIME is another XAI tool designed to interpret ML model predictions by locally approximating the model using a simple interpretable surrogate (Ribeiro et al., 2016). LIME generates an explanation for a specific prediction by perturbing the input, observing its effect on the output, and then fitting a simple linear model to local data. This produces a set of feature weights that indicate how each feature influences the prediction. Features with positive weights contributed positively, whereas those with negative weights had a negative influence on the prediction. The magnitude of a weight reflects the strength of the impact of a feature; thus, features can be ranked by their absolute weights. Unlike SHAP, which provides global and local insights, LIME focuses on local (instance-specific) explanations. In the study, LIME was employed to produce per-instance explanations, highlighting how particular operational

variables affect effluent quality predictions on individual days or under specific WWTP scenarios.

Using SHAP and LIME, a comprehensive understanding of model behavior is obtained. SHAP helped to consistently identify important predictors across the entire dataset (global importance), whereas LIME offered detailed case-specific explanations for individual predictions (local importance). Together, these XAI methods enhance model transparency and provide actionable insights for WWTP operational monitoring and management by revealing both the overall driving factors and the circumstances under which different factors become important.

4.2.5 ML models

RF is an ensemble learning technique and one of the most widely used ML methods in environmental modeling, owing to its robustness and ability to handle nonlinear relationships. It employs multiple decision trees to generate predictions (Tyralis et al., 2019; Jiang et al., 2023; Szomolányi and Clement, 2023; Sun et al., 2024). An important feature of this algorithm is its ability to provide insights into the importance of each variable, which is essential for further analysis (Zhang et al., 2021). In contrast, XGBoost is an optimized and highly efficient gradient boosting algorithm. XGBoost has become a popular choice for both regression and classification tasks, owing to its superior performance and scalability. This technique enhances the traditional gradient boosting by incorporating regularization techniques and parallel processing capabilities. XGBRegressor class from the XGBoost library was implemented to create an XGBoost model. RF and XGBoost were chosen for this study because of their widespread applications in the water sector. For the RF model, RandomForestRegressor class from the

sklearn.ensemble module was imported. The GridSearch method is used to determine the optimal hyperparameters for building the RF model. The tuning process involved varying the number of trees (n_estimators: 50, 100, 150, 200, 250, and 300), maximum depth of each tree (max_depth: None, 5, 10, 15, 20), minimum number of samples required to split an internal node (min_samples_split: 2, 5, 10, 15, 20), and minimum number of samples required in a leaf node (min_samples_leaf: 1, 2, 4, 6, 8). Additionally, the number of features considered for each split (max_features: 'auto,' 'sqrt,' 'log2') was fine-tuned to balance the model accuracy and the computational efficiency.

For XGBoost, the XGBRegressor class was imported from the XGBoost library. The tuning process involved several trees ranging from 50 to 300 to determine the optimal number of boosting rounds and learning rates from 0.01 to 0.3. To prevent overfitting and control tree complexity, the maximum depth was varied from 3 to 10, and the minimum sum of instance weights required in a child node varied from 1 to 5. Further optimization included tuning subsample from 0.6 to 1.0 to regulate the fraction of samples used per boosting round and subsample ratio of columns from 0.6 to 1.0 to manage the number of features considered in each tree. The optimal hyperparameter combination, identified through systematic tuning, enhanced model generalization while maintaining computational efficiency and ensured reliable predictions in wastewater treatment applications.

4.2.6 Model training and evaluation

For each WWTP dataset, the data were split into training, validation, and testing subsets in chronological order. The first 85% of the time-series data was used for model development, and the remaining 15% was used for the final test. This temporal split ensures that the test set

represents “future” data not seen by the model during training, mimicking a real-world forecasting scenario. From the initial 85% segment, the data were further divided into 70% for training (model fitting), and 15% for validation (model tuning). The model learns patterns and relationships from the training set, whereas the validation set is used to fine-tune the hyperparameters and prevent overfitting, without examining the test data. Reserving the independent test set until the end gives unbiased evaluation of how the model generalizes to the new data. K-fold cross-validation was performed to reduce the risk of overfitting (Zhang and Liu, 2023) by splitting the entire dataset into five equal sections. Thus, the 70% training set was further divided into five folds for cross-validation, where four folds were used for training and, 1-fold was used for validation. Using a well-known hyperparameter-tuning method, grid search, various hyperparameter combinations for the RF and XGBoost models were tested to determine the optimal set that yielded the best performance for certain validation metrics.

Four widely used assessment metrics were used to evaluate the performance of ML models: Mean Absolute Error (MAE), Mean Squared Error (MSE), R-squared (R^2), and Root Mean Squared Error (RMSE). MAE is a valuable metric that assesses the average magnitude of errors between predicted and actual values, offering insights into how closely predictions align with the actual outcomes. The MSE complements this by calculating the average squared difference between the predicted and actual values, thereby helping illustrate the potential magnitude of deviations in the predictions. R^2 is particularly useful because it quantifies the percentage of the variance in the data that the model can explain. With values ranging from 0 to 1, a value of 1 indicates a perfect explanation of variability, whereas a value of 0 suggests that the model does not capture any variance. The RMSE provides the average size of the

residuals and is always non-negative; lower values indicate a stronger fit to the data. Together, these metrics offer constructive feedback on the precision, goodness-of-fit, and accuracy of the ML models.

$$MAE = \frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{n} \quad (1)$$

$$MSE = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n} \quad (2)$$

$$R^2 = 1 - \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}} \quad (3)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2} \quad (4)$$

where $\hat{y}_i$ is the predicted value, y_i the experimental data, and n the number of test observations.

4.3 Results and discussion

4.3.1 FS selection

For small-scale facilities (Monroe and Sheboygan WWTPs), the feature selection methods yielded largely consistent rankings of important variables. In the Monroe WWTP, $CBOD_e$, TKN_e, and $Flow_e$ emerged as key predictors across all FS methods, with BOD_e and TSS_e also identified as influential in certain methods (Figure 4.5). In the Sheboygan WWTP, $CBOD_e$, $(NH_3\text{-}N)_e$, TSS_e, and BOD_i were repeatedly found to be important features (Figure 4.6). The inclusion of a BOD_i among the top predictors for Sheboygan suggests that both influent and effluent measurements can be significant drivers of effluent quality in smaller plants. Similar core predictors were observed in large-scale facilities. In the Madison WWTP, effluent $CBOD_e$,

$Flow_e$, and TSS_e were frequently ranked among the top features, along with influent measures such as $CBOD_i$ and TSS_i (Figure 4.7). Likewise, in the Milwaukee WWTP, $Flow_e$, $(NH_3\text{-}N)_e$, TSS_e, and TP_e consistently appeared as dominant predictors across the various FS methods (Figure 4.8). Overall, certain variables, particularly effluent BOD, flow, TP, and TSS, were repeatedly identified as significant predictors of the effluent quality across all plants. This agreement among different FS techniques and across multiple facilities highlights the critical role of these key variables in modeling WWTP effluent quality.

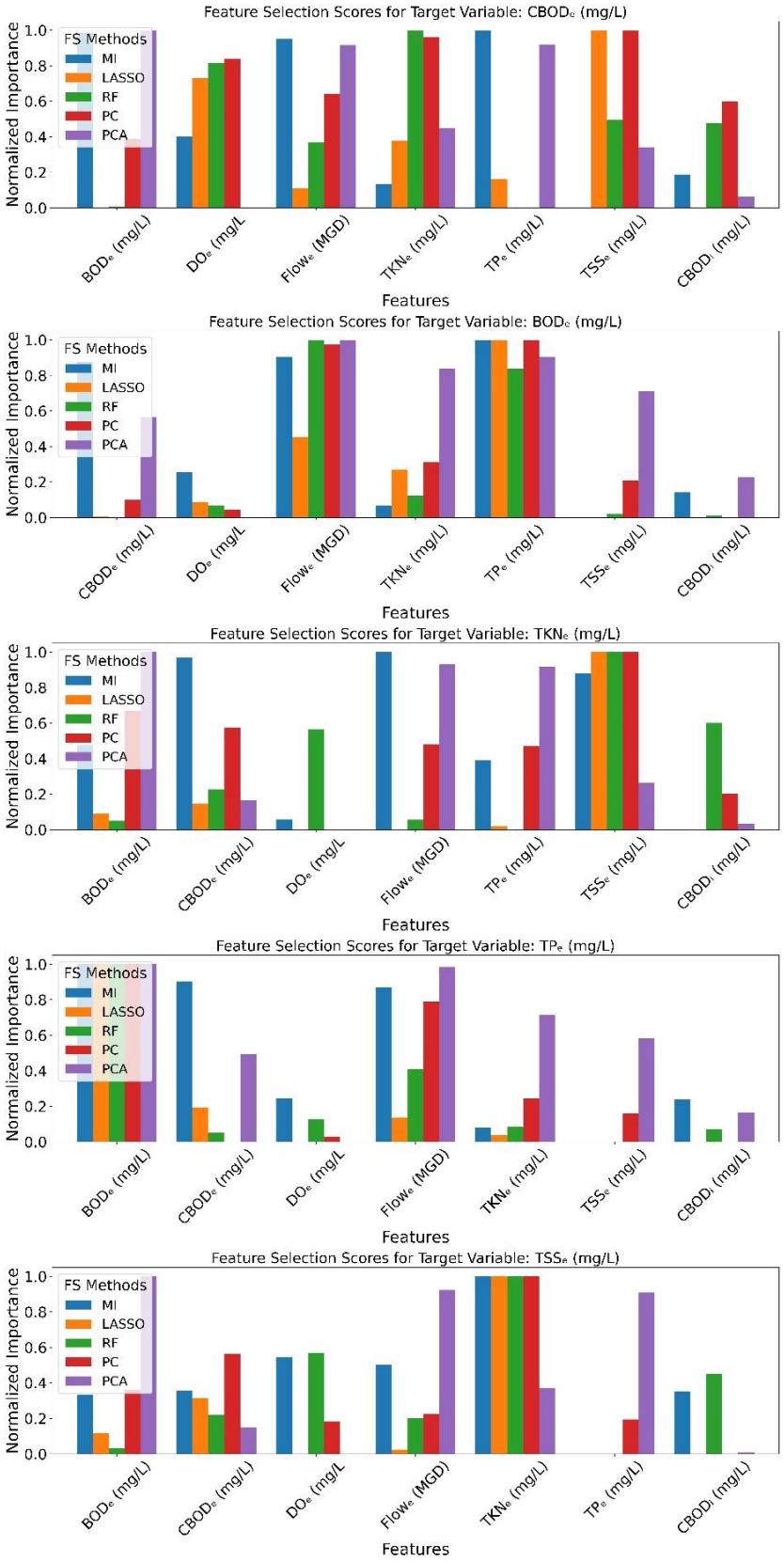

Figure 4.5. Feature selection scores of Monroe WWTP target variable

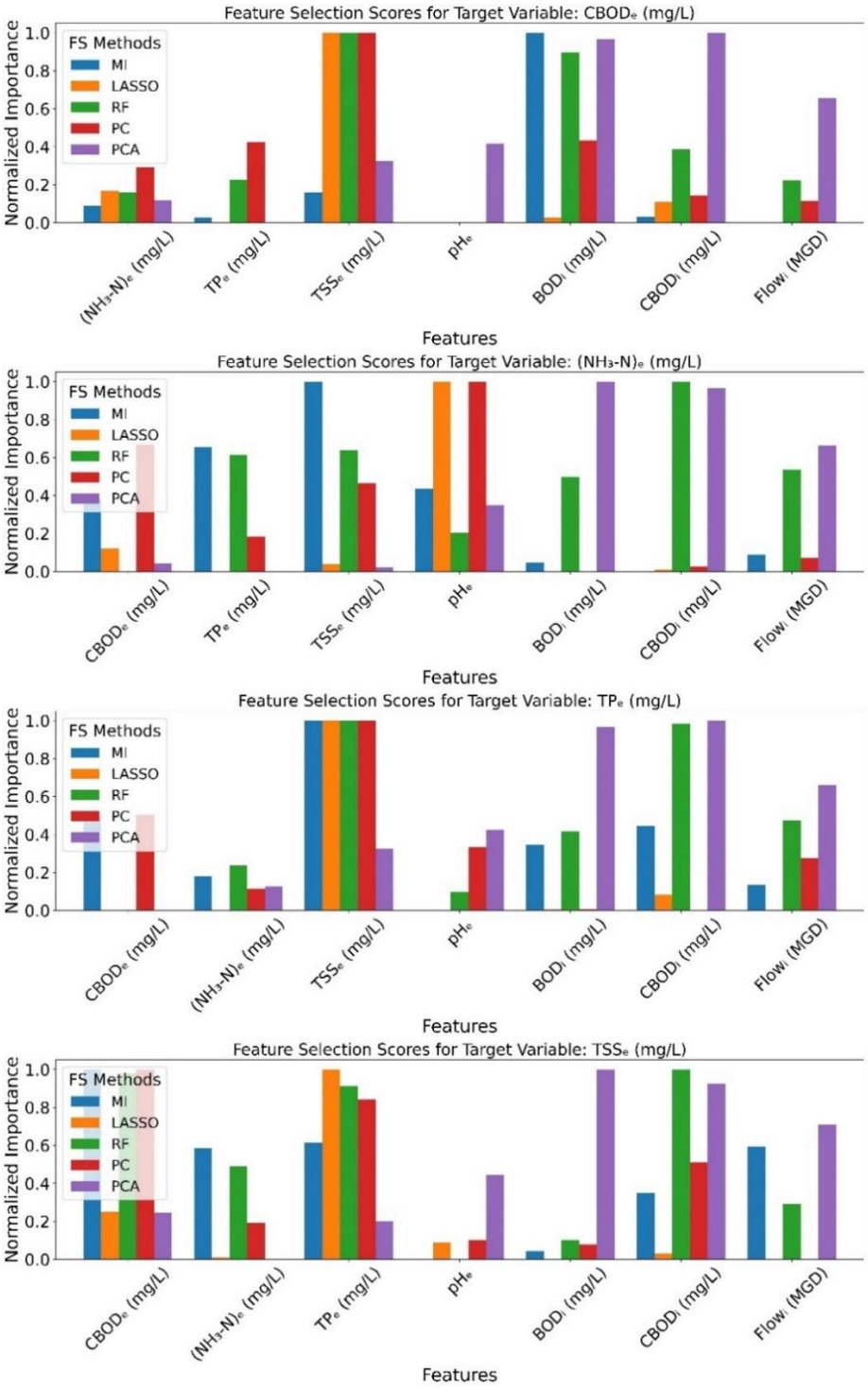

Figure 4.6. Feature selection of Sheboygan WWTP target variable

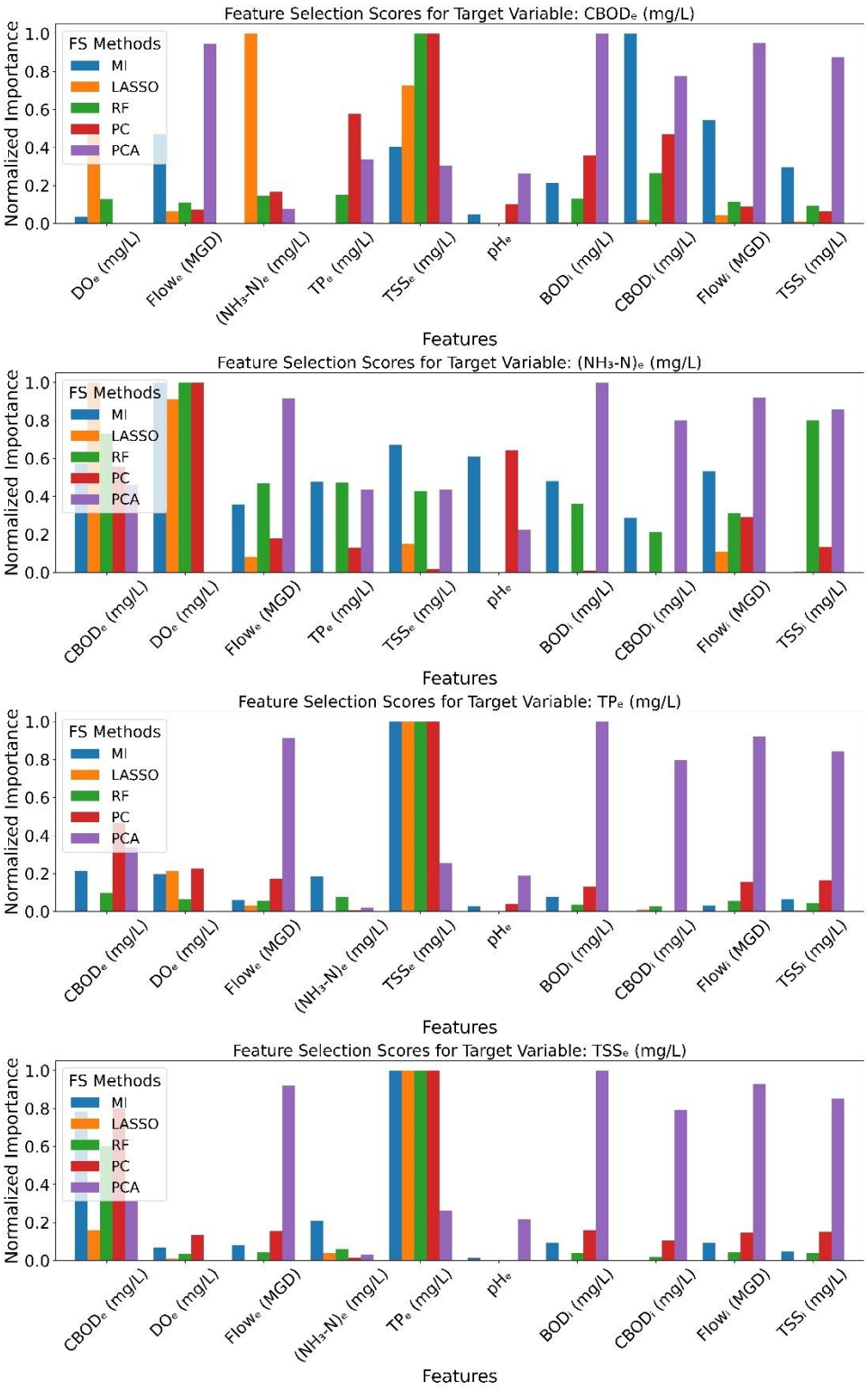

Figure 4.7. Feature selection scores of Madison WWTP target variable

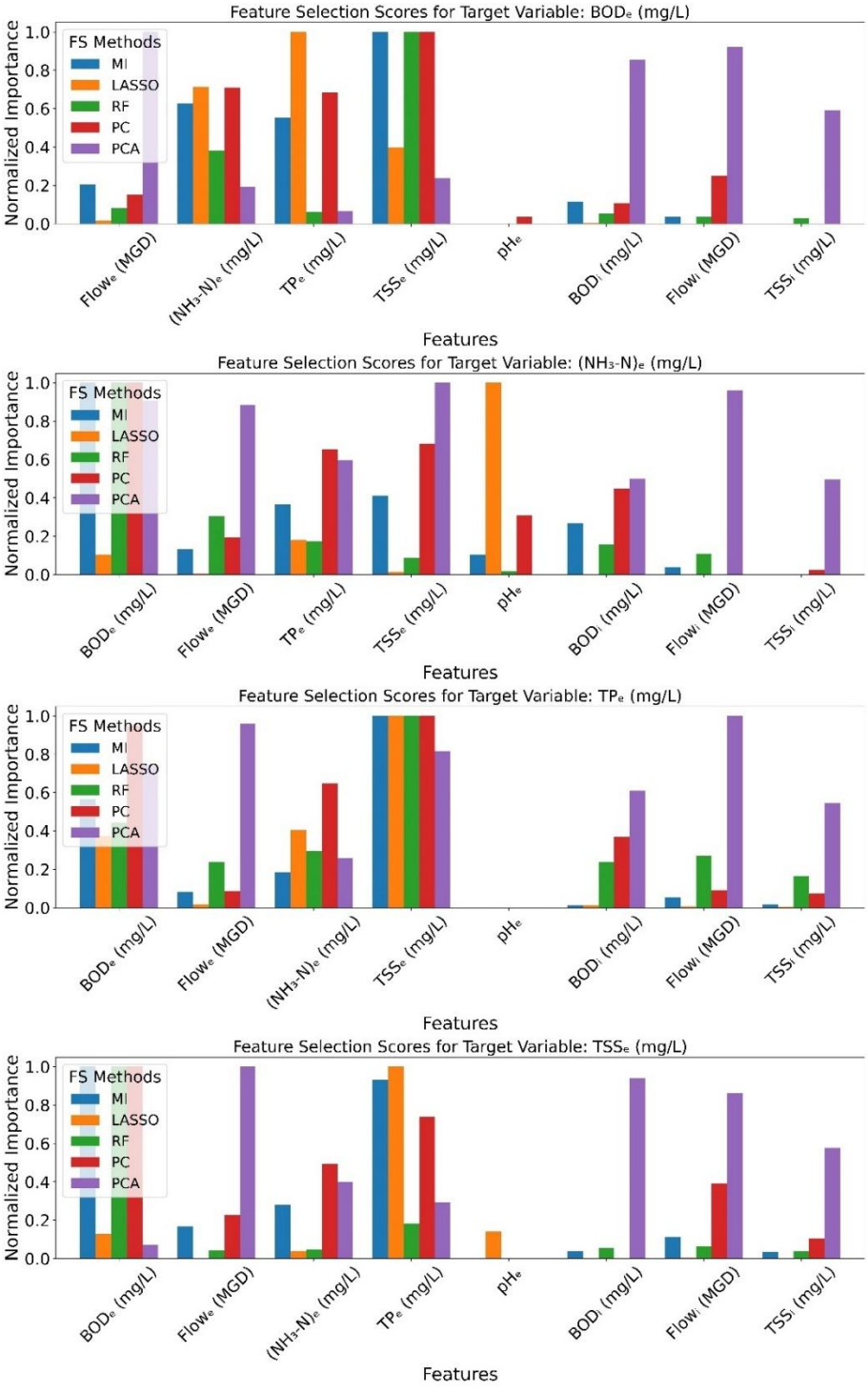

Figure 4.8. Feature selection scores of Milwaukee WWTP target variable

4.3.2 SHAP and LIME

The SHAP and LIME analyses provided critical insights into the prediction mechanisms of the ML models for all the target variables across the four WWTPs. Figures 4.9 and 4.10 illustrate how these tools explain the model predictions. Notably, the important predictors identified by SHAP and LIME varied somewhat between the target variables, reflecting the complex interplay of physical, chemical, and biological processes within the WWTPs.

For BOD_e predictions, both SHAP and LIME consistently identified $Flow_e$, $(NH_3\text{-}N)_e$, TP_e, and TSS_e as the dominant contributors in both the RF and XGBoost models. These variables represent major factors such as organic load, nutrient concentration, solid content, and hydraulic conditions, all of which significantly affect BOD_e levels. For $CBOD_e$, both SHAP and LIME analysis indicated that BOD_i, $Flow_e$, TKN_e, and TSS_e were among the top predictors across all WWTPs. These results imply that both the incoming organic load and nitrogen and the solids content are crucial for determining $CBOD_e$.

TSS_e predictions were strongly driven by other effluent quality variables according to SHAP. Both SHAP and LIME values indicated that BOD_e, $CBOD_e$, TKN_e, and TP_e had substantial impacts on the effluent TSS levels. This is intuitive because higher organic loads and nutrient concentrations were associated with changes in biomass and particulate matter during the treatment process. LIME, focusing on individual instances, sometimes ranked influential factors such as $CBOD_i$ and $Flow_i$ as influential for TSS_e. For TP_e, both SHAP and LIME were largely in agreement. They consistently highlighted the predictive value of BOD_e, $CBOD_e$, $(NH_3\text{-}N)_e$, and TSS_e.

SHAP results and LIME for TKN_e indicated that variables related to organic matter and solids were influential as BOD_e and $CBOD_e$ were frequently top contributors. This suggests that the effluent organic nitrogen and particulate-associated nitrogen might be linked to the overall organic load and solid retention in the system. LIME outputs for TKN_e also pointed to TP_e.

For $(NH_3\text{-}N)_e$, both XAI methods identified a mix of carbon, solids, and phosphorus indicators as key explanatory variables. SHAP and LIME analyses emphasized the roles of BOD_i, BOD_e, $CBOD_e$, TSS_e, and TP_e in the ammonia predictions, indicating that higher organic carbon and solids, as well as higher effluent phosphorus, were associated with changes in ammonia removal or production. This alignment between SHAP and LIME suggests strong coupling between carbon usage, solids, and phosphorus in the transformation of ammonia during the treatment process.

In summary, the SHAP and LIME results were largely complementary. SHAP was well-suited for identifying consistent, globally important drivers of model predictions across the entire dataset and for each target variable. LIME provides insight into instance-specific nuances, capturing how the influence of certain features can vary under different conditions or in different plants. The combination of both techniques confirms that the models are learning relationships that make physical and chemical sense and underscore the value of local explanations to detect when and why a model might rely on an unexpected factor for a particular prediction. These XAI insights can help WWTP operators and engineers trust ML model outputs and provide guidance on which variables are most critical for monitoring and controlling effluent quality.

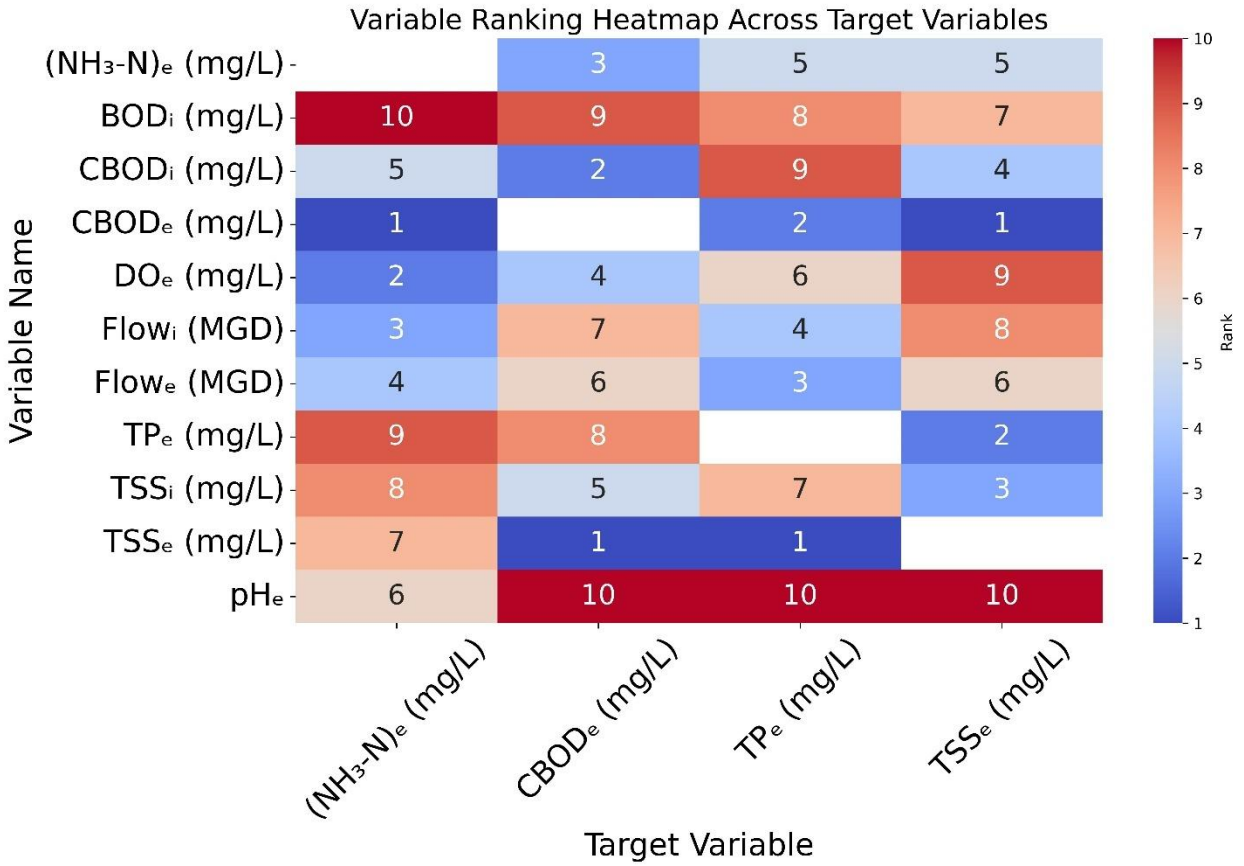

Figure 4.9. SHAP Madison for XGBoost model

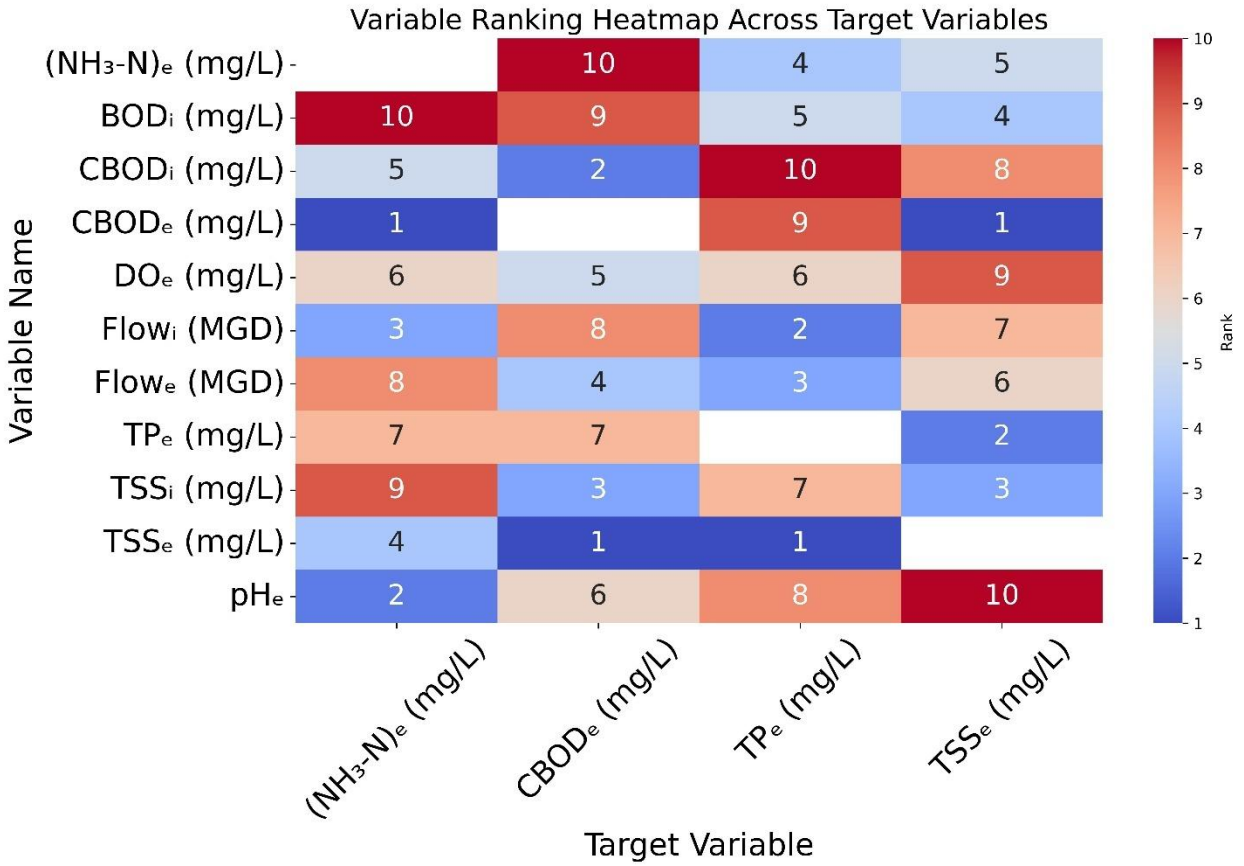

Figure 4.10. LIME Madison for XGBoost model

4.3.3 Model performance

The predictive performance of the RF and XGBoost models was evaluated for each WWTP using the metrics described earlier for the training, validation, and testing sets. Overall, both models fit the training data well, with high R^2 and low error metrics. However, their ability to generalize the test data varied notably between smaller and larger plants. In the Monroe WWTP (small-scale), the test set R^2 values for most target variables were negative (except for TP_e), indicating that the models struggled to make accurate predictions beyond the training period (Table 4.5). This poor test performance in Monroe suggests possible overfitting or significant shifts in the data during the test period, which the models could not capture. A

similar trend was observed for the Sheboygan WWTP. For instance, for $(NH_3\text{-}N)_e$ model showed a sharp drop in R^2 from training to testing (Table 4.6). The $(NH_3\text{-}N)_e$ test R^2 in Sheboygan was substantially lower (even negative) despite the high training R^2, likely due to the high variability in influent nitrogen loading and treatment efficiency, which was inconsistent with the training data patterns.

Table 4.5. Metrices for Monroe WWTP

WWTP name	ML model	Target Variable	Set	MAE	MSE	R^2	RMSE	Runtime (s)
MONROE	RF	BOD_e	Training	0.18	0.12	0.96	0.35	708.65
MONROE	RF	BOD_e	Validation	0.47	0.78	0.58	0.88	0.01
MONROE	RF	BOD_e	Test	0.54	0.63	-0.35	0.79	0.01
MONROE	RF	$CBOD_e$	Training	0.23	0.12	0.78	0.34	674.73
MONROE	RF	$CBOD_e$	Validation	0.40	0.34	0.40	0.58	0.01
MONROE	RF	$CBOD_e$	Test	0.55	0.63	-0.36	0.79	0.01
MONROE	RF	TP_e	Training	0.42	0.44	0.76	0.66	735.80
MONROE	RF	TP_e	Validation	0.46	0.61	0.39	0.78	0.00
MONROE	RF	TP_e	Test	0.53	0.90	0.09	0.95	0.02
MONROE	RF	TSS_e	Training	0.07	0.01	0.63	0.09	698.20
MONROE	RF	TSS_e	Validation	0.10	0.01	0.23	0.12	0.03
MONROE	RF	TSS_e	Test	0.07	0.01	-0.05	0.10	0.02
MONROE	XGBoost	BOD_e	Training	0.31	0.23	0.91	0.48	445.11
MONROE	XGBoost	BOD_e	Validation	0.47	0.71	0.61	0.84	0.00
MONROE	XGBoost	BOD_e	Test	0.52	0.57	-0.23	0.76	0.00
MONROE	XGBoost	$CBOD_e$	Training	0.27	0.15	0.72	0.39	581.58
MONROE	XGBoost	$CBOD_e$	Validation	0.38	0.31	0.45	0.56	0.00
MONROE	XGBoost	$CBOD_e$	Test	0.57	0.67	-0.45	0.82	0.02
MONROE	XGBoost	TP_e	Training	0.41	0.38	0.79	0.61	442.73
MONROE	XGBoost	TP_e	Validation	0.47	0.62	0.39	0.79	0.00
MONROE	XGBoost	TP_e	Test	0.52	0.88	0.11	0.94	0.00
MONROE	XGBoost	TSS_e	Training	0.07	0.01	0.60	0.09	557.57
MONROE	XGBoost	TSS_e	Validation	0.10	0.01	0.22	0.12	0.00
MONROE	XGBoost	TSS_e	Test	0.07	0.01	-0.06	0.10	0.00

Table 4.6. Metrices for Sheboygan WWTP

WWTP name	ML model	Target Variable	Set	MAE	MSE	R^2	RMSE	Runtime (s)
SHEBOYGAN	RF	$(NH_3\text{-}N)_e$	Training	0.72	1.22	0.70	1.10	699.83
SHEBOYGAN	RF	$(NH_3\text{-}N)_e$	Validation	1.08	2.82	0.25	1.68	0.04
SHEBOYGAN	RF	$(NH_3\text{-}N)_e$	Test	0.92	1.50	-0.74	1.22	0.02
SHEBOYGAN	RF	TP_e	Training	0.06	0.01	0.52	0.09	685.73
SHEBOYGAN	RF	TP_e	Validation	0.08	0.02	0.23	0.13	0.02
SHEBOYGAN	RF	TP_e	Test	0.09	0.02	0.18	0.13	0.00
SHEBOYGAN	RF	TSS_e	Training	0.41	0.28	0.90	0.53	702.59
SHEBOYGAN	RF	TSS_e	Validation	0.82	1.22	0.49	1.11	0.02
SHEBOYGAN	RF	TSS_e	Test	1.05	1.90	0.19	1.38	0.01
SHEBOYGAN	XGBoost	$(NH_3\text{-}N)_e$	Training	0.55	0.67	0.83	0.82	539.04
SHEBOYGAN	XGBoost	$(NH_3\text{-}N)_e$	Validation	1.06	2.68	0.29	1.64	0.00
SHEBOYGAN	XGBoost	$(NH_3\text{-}N)_e$	Test	0.92	1.56	-0.82	1.25	0.02
SHEBOYGAN	XGBoost	TP_e	Training	0.07	0.01	0.44	0.09	523.68
SHEBOYGAN	XGBoost	TP_e	Validation	0.08	0.02	0.26	0.13	0.00
SHEBOYGAN	XGBoost	TP_e	Test	0.09	0.02	0.17	0.13	0.02
SHEBOYGAN	XGBoost	TSS_e	Training	0.72	0.86	0.69	0.93	529.73
SHEBOYGAN	XGBoost	TSS_e	Validation	0.84	1.27	0.47	1.13	0.00
SHEBOYGAN	XGBoost	TSS_e	Test	1.05	1.89	0.19	1.38	0.00

Table 4.7. Metrices for Madison WWTP

WWTP name	ML model	Target Variable	Set	MAE	MSE	R^2	RMSE	Runtime (s)
MADISON	RF	$CBOD_e$	Training	0.13	0.05	0.91	0.22	893.30
MADISON	RF	$CBOD_e$	Validation	0.31	0.27	0.56	0.52	0.03
MADISON	RF	$CBOD_e$	Test	0.53	0.60	0.28	0.78	0.04
MADISON	RF	$(NH_3\text{-}N)_e$	Training	0.09	0.03	0.90	0.18	929.43
MADISON	RF	$(NH_3\text{-}N)_e$	Validation	0.22	0.14	-0.06	0.37	0.01
MADISON	RF	$(NH_3\text{-}N)_e$	Test	0.32	0.24	-0.25	0.49	0.00
MADISON	RF	TP_e	Training	0.03	0.00	0.89	0.05	931.93
MADISON	RF	TP_e	Validation	0.07	0.01	0.71	0.10	0.03
MADISON	RF	TP_e	Test	0.08	0.01	0.32	0.10	0.02
MADISON	RF	TSS_e	Training	0.33	0.41	0.95	0.64	903.66
MADISON	RF	TSS_e	Validation	0.83	2.06	0.80	1.43	0.02
MADISON	RF	TSS_e	Test	0.95	3.65	-0.47	1.91	0.02
MADISON	XGBoost	$CBOD_e$	Training	0.18	0.06	0.88	0.25	779.92
MADISON	XGBoost	$CBOD_e$	Validation	0.33	0.29	0.52	0.54	0.02
MADISON	XGBoost	$CBOD_e$	Test	0.56	0.64	0.23	0.80	0.00
MADISON	XGBoost	$(NH_3\text{-}N)_e$	Training	0.16	0.08	0.73	0.29	734.60
MADISON	XGBoost	$(NH_3\text{-}N)_e$	Validation	0.22	0.16	-0.25	0.41	0.00

MADISON	XGBoost	$(NH_3\text{-}N)_e$	Test	0.29	0.23	-0.21	0.48	0.00
MADISON	XGBoost	TP_e	Training	0.05	0.01	0.79	0.07	718.03
MADISON	XGBoost	TP_e	Validation	0.07	0.01	0.66	0.11	0.02
MADISON	XGBoost	TP_e	Test	0.08	0.01	0.33	0.10	0.00
MADISON	XGBoost	TSS_e	Training	0.31	0.17	0.98	0.41	686.05
MADISON	XGBoost	TSS_e	Validation	0.81	1.82	0.82	1.35	0.01
MADISON	XGBoost	TSS_e	Test	0.95	5.43	-1.19	2.33	0.00

Table 4.8. Metrices for Milwaukee WWTP

WWTP name	ML model	Target Variable	Set	MAE	MSE	R^2	RMSE	Runtime (s)
MILWAUKEE	RF	BOD_e	Training	0.92	1.60	0.97	1.26	786.96
MILWAUKEE	RF	BOD_e	Validation	2.22	9.93	0.69	3.15	0.04
MILWAUKEE	RF	BOD_e	Test	3.40	18.94	0.62	4.35	0.03
MILWAUKEE	RF	$(NH_3\text{-}N)_e$	Training	0.36	0.53	0.80	0.73	791.94
MILWAUKEE	RF	$(NH_3\text{-}N)_e$	Validation	0.52	0.80	0.40	0.90	0.01
MILWAUKEE	RF	$(NH_3\text{-}N)_e$	Test	1.00	3.08	0.25	1.76	0.02
MILWAUKEE	RF	TP_e	Training	0.11	0.04	0.63	0.20	793.77
MILWAUKEE	RF	TP_e	Validation	0.14	0.04	0.38	0.20	0.00
MILWAUKEE	RF	TP_e	Test	0.16	0.04	0.41	0.21	0.02
MILWAUKEE	RF	TSS_e	Training	0.87	1.96	0.95	1.40	789.63
MILWAUKEE	RF	TSS_e	Validation	1.90	7.85	0.61	2.80	0.01
MILWAUKEE	RF	TSS_e	Test	1.94	6.91	0.69	2.63	0.00
MILWAUKEE	XGBoost	BOD_e	Training	1.76	5.35	0.89	2.31	572.61
MILWAUKEE	XGBoost	BOD_e	Validation	2.17	9.79	0.70	3.13	0.00
MILWAUKEE	XGBoost	BOD_e	Test	3.20	16.40	0.67	4.05	0.02
MILWAUKEE	XGBoost	$(NH_3\text{-}N)_e$	Training	0.46	0.63	0.76	0.80	489.31
MILWAUKEE	XGBoost	$(NH_3\text{-}N)_e$	Validation	0.50	0.75	0.43	0.87	0.01
MILWAUKEE	XGBoost	$(NH_3\text{-}N)_e$	Test	1.03	3.20	0.22	1.79	0.00
MILWAUKEE	XGBoost	TP_e	Training	0.12	0.04	0.65	0.20	561.53
MILWAUKEE	XGBoost	TP_e	Validation	0.14	0.04	0.35	0.20	0.01
MILWAUKEE	XGBoost	TP_e	Test	0.16	0.04	0.41	0.21	0.00
MILWAUKEE	XGBoost	TSS_e	Training	0.78	1.11	0.97	1.06	506.23
MILWAUKEE	XGBoost	TSS_e	Validation	1.85	7.08	0.65	2.66	0.02
MILWAUKEE	XGBoost	TSS_e	Test	2.09	7.89	0.64	2.81	0.00

In Madison WWTP (large-scale), the models showed more robust performance. For example, the predictions of $CBOD_e$ maintained reasonably good accuracy from training to testing (Table 4.7), indicating that the models captured stable relationships for CBOD in that

plant. However, even in Madison, the (NH_3-N)$_e$) model yielded a negative R^2 on the test set, underscoring the difficulty of modeling nitrogen transformations that can be influenced by complex microbial dynamics and occasional process upsets. On the other hand, TSS_e and TP_e predictions in Madison had relatively better test performance, although still lower than the training performance as expected. Among the four plants, Milwaukee WWTP (the largest facility) had the strongest predictive results. For effluent BOD_e in Milwaukee, the RF model achieved an R^2 of about 0.97 on training and 0.62 on the test set, while the XGBoost model achieved a comparable test R^2 of 0.67 (Table 4.8). The smaller gap between the training and testing performances in Milwaukee indicates better generalization. TSS_e predictions in Milwaukee were also relatively strong for both RF and XGBoost, maintaining high accuracy into the test period. These results suggest that the models trained on large-scale Milwaukee plant data were able to capture the underlying patterns more reliably, likely because of the greater volume of data and more consistent operation of the facility. Another important finding was the difference in model transferability across plants at different scales. When a model trained on a large-scale WWTP dataset was applied to predict another large-scale WWTP, it generalized reasonably well. However, the models trained using large-scale data struggled significantly when tested on small-scale WWTP data. For example, a model developed on Milwaukee WWTP data achieved relatively strong performance in predicting Madison WWTP's TP_e (R^2 value 0.44 using RF and 0.25 using XGBoost), whereas performance sharply deteriorated when predicting small-scale plants like Monroe. Conversely, models trained on small-scale plant data (Monroe or Sheboygan) exhibited substantial declines in accuracy when applied to large-scale facilities (Milwaukee or Madison), often resulting in highly negative R^2 values. Several factors contribute

to scale-dependent generalization. One primary issue is the significant difference in the statistical distributions and variability of influent and effluent variables between large and small WWTPs. Large plants generally maintain steadier operational conditions and richer datasets, thereby allowing the models to identify more stable and consistent patterns. Small plants experience greater variability in their influent characteristics, limited data availability, and frequent operational disturbances. Furthermore, only the common input variables between plants were used during these cross-plant evaluations, resulting in fewer available predictors and potentially poorer model performances. This reduced feature set likely exacerbated the poor transferability, especially from larger to smaller plants, because fewer variables decreased the flexibility and accuracy of the models. Therefore, this discrepancy in performance emphasizes the importance of scale-specific model calibration and suggests the potential benefits of transfer learning techniques in improving model adaptability across different plant scales.

4.4 Conclusions

This study assessed the performance of two ML models, RF and XGBoost, in predicting key effluent quality variables at WWTPs with varying capacities. FS and XAI tools (SHAP and LIME) identified and interpreted the influence of the input variables on the model predictions. The results demonstrated that both the FS and XAI tools provided consistent and interpretable insights into variable importance across all WWTPs, regardless of scale. In particular, the models and XAI analyses consistently highlighted crucial factors such as organic load, nutrient levels, and solids as primary drivers of effluent quality. These tools have proven to be valuable

for enhancing the transparency of ML models and identifying key operational parameters that affect effluent outcomes. However, the predictive accuracy of ML models varied significantly among facilities of different sizes. Large-scale WWTPs (e.g., Madison and Milwaukee) exhibited more stable and reliable model performance, likely owing to their more consistent operations and the availability of larger, less noisy datasets. In contrast, the small-scale WWTPs (Monroe and Sheboygan) experienced reduced prediction accuracy, which can be attributed to limited data, higher relative noise, greater variability in influent characteristics, and less consistent operational practices. Future research should focus on developing scalable, data-efficient ML frameworks that can be adapted to different plant sizes. Overall, such initiatives are important for advancing the practical integration of interpretable AI into real-world wastewater management, ultimately contributing to more efficient and sustainable WWTP operation.

4.5 References

1. Aghdam, E., Mohandes, S. R., Manu, P., Cheung, C., Yunusa-Kaltungo, A., & Zayed, T. (2023). Predicting quality parameters of wastewater treatment plants using artificial intelligence techniques. Journal of Cleaner Production, 405, 137019.

2. Alvi, M., Batstone, D., Mbamba, C. K., Keymer, P., French, T., Ward, A., ... & Cardell-Oliver, R. (2023). Deep learning in wastewater treatment: a critical review. Water Research, 245, 120518.

3. Arismendy, L., Cárdenas, C., Gómez, D., Maturana, A., Mejía, R., & Quintero M, C. G. (2020). Intelligent system for the predictive analysis of an industrial wastewater treatment process. Sustainability, 12(16), 6348.

4. Cechinel, M. A. P., Neves, J., Fuck, J. V. R., de Andrade, R. C., Spogis, N., Riella, H. G., ... & Soares, C. (2024). Enhancing wastewater treatment efficiency through machine learning-driven effluent quality prediction: A plant-level analysis. Journal of Water Process Engineering, 58, 104758.

5 El-Rawy, M., Abd-Ellah, M. K., Fathi, H., & Ahmed, A. K. A. (2021). Forecasting effluent and performance of wastewater treatment plant using different machine learning techniques. Journal of Water Process Engineering, 44, 102380.

6. Hu, Y., Wei, R., Yu, K., Liu, Z., Zhou, Q., Zhang, M., ... & Qu, S. (2024). Exploring sludge yield patterns through interpretable machine learning models in China's municipal wastewater treatment plants. Resources, Conservation and Recycling, 204, 107467.

7. Huang, X., Kroening, D., Ruan, W., Sharp, J., Sun, Y., Thamo, E., ... & Yi, X. (2020). A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. Computer Science Review, 37, 100270.

8. Jiang, M., Wang, J., Hu, L., & He, Z. (2023). Random forest clustering for discrete sequences. Pattern Recognition Letters, 174, 145-151.

9. Li, R., Feng, K., An, T., Cheng, P., Wei, L., Zhao, Z., ... & Zhu, L. (2024). Enhanced insights into effluent prediction in wastewater treatment plants: Comprehensive deep learning model explanation based on shap. ACS ES&T Water, 4(4), 1904-1915.

10. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

11. Newhart, K. B., Klanderman, M. C., Hering, A. S., & Cath, T. Y. (2023). A holistic evaluation of multivariate statistical process monitoring in a biological and membrane treatment system. ACS Es&t Water, 4(3), 913-924.

12. Park, J., Lee, W. H., Kim, K. T., Park, C. Y., Lee, S., & Heo, T. Y. (2022). Interpretation of ensemble learning to predict water quality using explainable artificial intelligence. Science of the Total Environment, 832, 155070.

13. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).

14. Shao, S., Fu, D., Yang, T., Mu, H., Gao, Q., & Zhang, Y. (2023). Analysis of machine learning models for wastewater treatment plant sludge output prediction. Sustainability, 15(18), 13380.

15. Sun, Z., Wang, G., Li, P., Wang, H., Zhang, M., & Liang, X. (2024). An improved random forest based on the classification accuracy and correlation measurement of decision trees. Expert Systems with Applications, 237, 121549.

16. Shyu, H. Y., Castro, C. J., Bair, R. A., Lu, Q., & Yeh, D. H. (2023). Development of a soft sensor using machine learning algorithms for predicting the water quality of an onsite wastewater treatment system. ACS Environmental Au, 3(5), 308-318.

17. Szomolányi, O., & Clement, A. (2023). Use of random forest for assessing the effect of water quality parameters on the biological status of surface waters. GEM-International Journal on Geomathematics, 14(1), 20.

18. Tyralis, H., Papacharalampous, G., & Langousis, A. (2019). A brief review of random forests for water scientists and practitioners and their recent history in water resources. Water, 11(5), 910.

19. Wang, D., Thunéll, S., Lindberg, U., Jiang, L., Trygg, J., Tysklind, M., & Souihi, N. (2021). A machine learning framework to improve effluent quality control in wastewater treatment plants. Science of the total environment, 784, 147138.

20. Wei, X., Yu, J., Tian, Y., Ben, Y., Cai, Z., & Zheng, C. (2023). Comparative performance of three machine learning models in predicting influent flow rates and nutrient loads at wastewater treatment plants. ACS ES&T Water, 4(3), 1024-1035.

21. Xu, Y., Wang, Z., Nairat, S., Zhou, J., & He, Z. (2023). Artificial intelligence-assisted prediction of effluent phosphorus in a full-scale wastewater treatment plant with missing phosphorus input and removal data. ACS ES&T Water, 4(3), 880-889.

22. Yu, J., Tian, Y., Jing, H., Sun, T., Wang, X., Andrews, C. B., & Zheng, C. (2023). Predicting regional wastewater treatment plant discharges using machine learning and population migration big data. ACS ES&T Water, 3(5), 1314-1328.

23. Zhang, S., Wang, H., & Keller, A. A. (2021). Novel machine learning-based energy consumption model of wastewater treatment plants. ACS ES&T Water, 1(12), 2531-2540.

24. Zhu, J. J., Borzooei, S., Sun, J., & Ren, Z. J. (2022). Deep learning optimization for soft sensing of hard-to-measure wastewater key variables. ACS ES&T Engineering, 2(7), 1341-1355.

CHAPTER 5: CONCLUSION

ML models provide less information on which variables are truly driving the model's predictions and to what extent they are contributing. The inability to explain variable importance poses significant challenges that limit trust in the model's results and hinder the ability to improve the model. Rationale behind ML predictions, the basis for trust in these predictions, and methods for error correction are some of the concerns, especially relevant in WWTP, where the reliability of ML practices is critical. This study serves as a practical demonstration of leveraging ML models and XAI to tackle real-world challenges in wastewater treatment, offering valuable insights for future endeavors in predictive modeling and process optimization.

In Chapter 1, the goals of the dissertation were discussed.

Chapter 2 presents advancements in predictive modeling for sludge production in WWTP, utilizing ML alongside XAI techniques to enhance accuracy and interpretability. The findings demonstrated the efficacy of GBM and GBT in predicting sludge production. While GBM exhibited superior performance, the study emphasizes the significance of selecting appropriate input variables to capture intricate relationships and ensure robust predictions on unseen data. XAI techniques, SHAP and LIME, identified key drivers of sludge production. This study serves as a practical demonstration of leveraging ML models and XAI to tackle real-world challenges in wastewater treatment, offering valuable insights for future endeavors in predictive modeling and process optimization.

Chapter 3 compared several XAI tools in predicting key WWTP variables using various ML models. ML models, ANN, GBM, XGBoost, and RF-GBM consistently outperform the others,

exhibiting strong prediction abilities with reduced errors and higher R^2 values. The use of SHAP and LIME enhances the interpretability of ML models by providing the impact of input variables on the model outputs. The reliability of XAI tools in identifying important WWTP factors is supported by the agreement of results between FS approaches and XAI tools.

Chapter 4 assessed the performance of two ML models, RF and XGBoost, in predicting key effluent quality variables at WWTPs with varying capacities. The results demonstrated that both the FS and XAI tools provided consistent and interpretable insights into variable importance across all WWTPs, regardless of scale. In particular, the models and XAI analyses consistently highlighted crucial factors such as organic load, nutrient levels, and solids as primary drivers of effluent quality. These tools have proven to be valuable for enhancing the transparency of ML models and identifying key operational parameters that affect effluent outcomes. However, the predictive accuracy of ML models varied significantly among facilities of different sizes. Large-scale WWTPs exhibited more stable and reliable model performance, likely owing to their more consistent operations and the availability of larger, less noisy datasets. In contrast, the small-scale WWTPs experienced reduced prediction accuracy, which can be attributed to limited data, higher relative noise, greater variability in influent characteristics, and less consistent operational practices. These findings underscore the need for tailored data strategies and enhanced monitoring in smaller facilities to optimize the effectiveness of ML applications.

In summary, the study explored the performance of various ML models across multiple WWTPs in the US. FS and XAI were used to test their effectiveness in understanding the influence of input variables on target variables. The results revealed that FS methods and XAI

tools were consistent regardless of the scale of WWTP. The tools were useful in picking influential variables regardless of their lack of knowledge of input variables on target variables from the real world.